DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN ELECTRÓNICA Y TELECOMUNICACIONES

TEMA: DESARROLLO DE UNA METODOLOGÍA BASADA EN MODELOS PARA EL RECONOCIMIENTO DE LOS NIVELES DE ESTRÉS EN EL DISCURSO DE UNA PERSONA EMPLEANDO PROCESAMIENTO DE LENGUAJE NATURAL.

AUTOR: LLUMIQUINGA GUAMÁN, LUIS ARMANDO

DIRECTOR: ALULEMA FLORES, DARWIN OMAR

SANGOLQUÍ

2017

CERTIFICACIÓN

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

CERTIFICACIÓN

Certifico que el trabajo de titulación, **"DESARROLLO DE UNA METODOLOGÍA BASADA EN MODELOS PARA EL RECONOCIMIENTO DE LOS NIVELES DE ESTRÉS EN EL DISCURSO DE UNA PERSONA EMPLEANDO PROCESAMIENTO DE LENGUAJE NATURAL"** realizada por el señor **LUIS ARMANDO LLUMIQUINGA GUAMÁN**, ha sido revisado en su totalidad y analizado por el software antiplagio, el mismo cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas – ESPE, por lo tanto me permito acreditarlo y autorizar al señor **LUIS ARMANDO LLUMIQUINGA GUAMAN** para que lo sustente públicamente.

Sangolquí, 18 de agosto del 2017

DARWIN OMAR ALULEMA FLORES

DIRECTOR

AUTORÍA DE RESPONSABILIDAD

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORÍA DE RESPONSABILIDAD

Yo, **LUIS ARMANDO LLUMIQUINGA GUAMAN**, con cédula de identidad Nº 172018691-3 declaro que este trabajo de titulación **"DESARROLLO DE UNA METODOLOGÍA BASADA EN MODELOS PARA EL RECONOCIMIENTO DE LOS NIVELES DE ESTRÉS EN EL DISCURSO DE UNA PERSONA EMPLEANDO PROCESAMIENTO DE LENGUAJE NATURAL"** ha sido desarrollada considerando los métodos de investigación existentes, así como también se ha respetado los derechos intelectuales de terceros considerándose en las citas bibliográficas.

Consecuentemente declaro que es trabajo de mi autoría, en virtud de ello me declaro responsable del contenido, veracidad y alcance de la investigación mencionada.

Sangolquí, 6 de septiembre del 2017

LUIS ARMANDO LLUMIQUINGA GUAMÁN

C.C.: 172018691-3

AUTORIZACIÓN

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORIZACIÓN

Yo, **LUIS ARMANDO LLUMIQUINGA GUAMAN**, autorizo a la Universidad de las Fuerzas Armadas – ESPE publicar en la biblioteca virtual de la institución el presente trabajo de titulación **"DESARROLLO DE UNA METODOLOGÍA BASADA EN MODELOS PARA EL RECONOCIMIENTO DE LOS NIVELES DE ESTRÉS EN EL DISCURSO DE UNA PERSONA EMPLEANDO PROCESAMIENTO DE LENGUAJE NATURAL"** cuyo contenido, ideas y criterios son de mi autoría y responsabilidad.

Sangolquí, 6 de septiembre del 2017

LUIS ARMANDO LLUMIQUINGA GUAMÁN

C.C: 172018691-3

DEDICATORIA

A mis hijas Danna y Salomé por ser mi razón para seguir luchando a pesar de las dificultades y por brindarme su cariño incondicional en todo este tiempo y a mis padres por ayudarme siempre a cumplir con las metas que me he propuesto.

Luis Armando Llumiquinga Guamán

AGRADECIMIENTO

Agradezco principalmente a Dios por permitirme culminar esta etapa de mi vida y por todas sus bendiciones hacía mí.

También quiero agradecer a mis padres Luz María y José Manuel por su ejemplo y por su apoyo moral durante toda mi etapa de formación profesional a mi esposa Gabriela por estar junto a mí apoyándome en mi carrera y a mis hermanos Jairo, Karina y Samanta.

A mi director de tesis Ing. Darwin Alulema por darme la oportunidad de demostrar mis habilidades en el desarrollo de este proyecto de tesis y por todos los conocimientos impartidos en las aulas de clase.

Luis Armando Llumiquinga Guamán

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El Procesamiento del Lenguaje Natural es una de las ramas de la inteligencia artificial y constituye una de las principales herramientas para el análisis de opiniones, sentimientos e ideas expresadas por el ser humano mediante el lenguaje escrito en cualquier idioma. El sistema desarrollado permite recolectar tweets de diferentes regiones del Ecuador para su posterior análisis, utiliza un clasificador de texto basado en el algoritmo de decisión Naive Bayes el cual predice si un mensaje contiene o no estrés haciendo uso de su entrenador que posee toda la información necesaria acerca del estrés que puede ser actualizada según la necesidad. Además, se probó el funcionamiento del sistema de detección del nivel de estrés en el análisis de varias evaluaciones del test psicológico de Sacks realizadas a un grupo de personas anónimas de diferentes edades. Este proyecto ofrece una metodología para la detección del estrés que es uno de los problemas sociales de mayor riesgo y que afectan en la actualidad a gran parte de la población ecuatoriana.

PALABRAS CLAVE:

- **PROCESAMIENTO DEL LENGUAJE NATURAL**
- **NLTK**
- **APRENDIZAJE DE MAQUINA**
- **TWITTER**
- **NAIVE BAYES**

ABSTRACT

The Natural Language Processing is one of the branches of artificial intelligence and constitutes one of the main tools for opinion analysis, feelings and ideas expressed by the human being through written language in any language. The developed system allows to collect tweets from different regions of the Ecuador for later analysis, uses a text classifier based on the Naive Bayes decision algorithm which predicts whether a message contains or not stress; using its trainer who has all the necessary information about stress that can be updated as needed. In addition, we tested the operation of the stress level detection system in the analysis of several assessments of the Sacks psychological test performed to a group of anonymous different ages people. This project offers a methodology for stress detection, that is one of the social problems of greater risk and that currently affect a large part of the Ecuadorian population.

KEYWORDS:

- **NATURAL LANGUAGE PROCESSING**
- **NLTK**
- **MACHINE LEARNING**
- **TWITTER**
- **NAIVE BAYES**

CAPÍTULO 1

1.1. Antecedentes

Existen diferentes modos y escalas de medición del estrés, siendo su enumeración completa difícil de relatar ya que es demasiado extensa, por mencionar algunos se tiene: Escalas de Apreciación del Estrés (Seara & Robles, 2017), Escala de percepción de estrés (Sender, Valles, Puig, Salamero, & Valdés, 2004), Test de estrés simple y de la tensión (Huaquin, Moyano, & Loaiza, 2000).

En el caso de los instrumentos de medición, se han realizado investigaciones que centran su atención en las reacciones fisiológicas (Pérez, De Macedo, Canelones, & Castés, 2002),utilizan cuestionarios (Macías, 2005), inventarios (Hernández, Polo, & Pozo, 1996) o escalas (Viñas & Caparrós, 2000), pero aún no se ha desarrollado una metodología que permita medir el estrés en las redes sociales mediante el uso de NLP.

El estudio de las fuentes más importantes que se han utilizado del estudio del síndrome del estrés, ofrece una clasificación de las clases de estresores existentes:

- Circunstancias que obligan a procesar información de forma rápida
- Estimulaciones del ambiente dañinas
- Sensaciones desagradables
- Cambios fisiológicos de la salud
- Incomunicación y encierro
- Asaltos del interés propio
- Imposición de un grupo
- Impotencia

Este trabajo de investigación se enfoca en las situaciones sociales más relevantes que pueden generar estrés, de acuerdo a los estudios realizados en el área de la psicología, que clasifican a las fuentes de estrés o estresores en varias categorías y se basa en la escala de estrés de (Holmes & Rahe, 1967) y en el estudio del DASS (Henry & Crawford, The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample, 2005). Esta escala mide los acontecimientos vitales más frecuentes y los valora según el grado de estrés que puede causar a las personas.

Esta escala de estrés está orientada a medir los acontecimientos vitales del último año e intentar valorar el riesgo de enfermedad debido al estrés. Para ello mide el estrés en unidades de estrés subjetivas que van de 0 a 100. Debido a la extensa cantidad de

categorías se analizarán únicamente las fuentes que puedan ser evaluadas mediante textos de acuerdo a la metodología planteada en este proyecto de tesis.

Tabla 1
Fuentes de estrés

FUENTE DE ESTRÉS (ESTRESOR)	PORCENTAJE DE ESTRES
Divorcio	73%
Muerte de un familiar cercano	63%
Lesión o enfermedad personal	53%
Despido del trabajo	47%
Enfermedad de un pariente cercano	44%
Cambio de la situación económica	38%
Discusiones con la pareja	35%
Cambio en las condiciones de vida	25%
Cambio de hábito de dormir	16%

Fuente: (Izarraga & Serra, 2016)

El gran crecimiento de los medios sociales como los blogs, o las redes sociales (Flores Vivar, 2009) han permitido a los usuarios expresar sus ideas, hacer comentarios, dar valoraciones y otras formas de expresión siendo este un medio bastante útil para el desarrollo de sistemas de análisis mediante NLP, ya que los usuarios inconscientemente plasman sus ansiedades, prejuicios, miedos.

1.2. Justificación e Importancia

El plan nacional del buen vivir 2013-2017 en el Ecuador tiene como uno de sus objetivos asegurar la soberanía y eficiencia de los sectores estratégicos para la transformación industrial y tecnológica, promueve en el inciso 11.3 literal g: "Establecer mecanismos de transferencia de tecnología en la normativa de telecomunicaciones, para permitir el desarrollo local de nuevas aplicaciones y servicios" (Senplades, 2013).

La Promoción de la Salud es una prioridad constitucional que se está operacionalizando en los planes, programas y proyectos de desarrollo local y nacional. Este nivel busca consolidarse dentro de una perspectiva renovada de abordaje de la salud, en un marco de descentralización, responsabilidad social, participación,

articulación intersectorial y empoderamiento de los diversos sujetos, el Ministerio de Salud Pública (MSP) del Ecuador cuenta con la Dirección Nacional de Promoción y Atención Integral, en la que se ha estructurado una unidad funcional para impulsar la promoción de la salud en el sector (Donoso, Herrera, & Aguinaga, 2004).

El estrés es un problema que afecta a la población del Ecuador debido a diversas causas como el desempleo, la crisis económica, los desastres naturales, condiciones de vida entre otras. Las enfermedades mentales más comunes del Ecuador son las derivadas de las condiciones de estrés a las que están sometidas diariamente las personas. El Instituto Nacional de Estadística y Censos (INEC) presentó la Encuesta de Condiciones de Vida que reflejan algunas de las causas que producen estrés en la población (INEC, 2014).

La detección del estrés en una persona puede advertir de futuras afecciones en su salud física y mental. El estrés también puede ocasionar conflictos con otras personas de su alrededor (Trucco, 2002).

En la actualidad no se ha desarrollado ninguna metodología que permita medir el nivel de estrés de una persona a partir de textos o discursos escritos obtenidos de las redes sociales. Por esta razón surge la necesidad de contar con una herramienta de diagnóstico que evidencie el nivel de estrés de una persona mediante el análisis de la información proporcionada en sus comentarios en la red social Twitter. El sistema propuesto para el reconocimiento del estrés también contribuirá al tratamiento de los trastornos emocionales surgidos en diferentes ámbitos de la sociedad.

1.3. Alcance

Este trabajo de investigación consiste en el desarrollo de un sistema basado en NLP, para el reconocimiento de los niveles de estrés que experimenta una persona analizando las expresiones publicadas en las redes sociales, el sistema se basa en el modelo de la escala de depresión ansiedad y estrés DASS-21 (Henry & Crawford, 2005), para el desarrollo de este sistema se utilizó el lenguaje programación Python que permite el uso de varias herramientas de NLP, tareas de procesamiento, muestras de corpus entre otros. Los tipos de datos incluyen fichas, etiquetas, trozos, árboles y características estructuradas.

Se realizó un estudio del estado del arte de las técnicas y algoritmos usados en el NLP y de ingeniería de modelos, así como de las herramientas de software para su desarrollo. La metodología de desarrollo del sistema se basó en la creación de modelos que permiten la representación abstracta de las características que definen el estrés.

Se desarrolló un sistema capaz de detectar los niveles de estrés a partir del texto escrito de una persona obtenido de la red social Twitter utilizando NLP con el fin de determinar el contexto y sentido del corpus del texto. El sistema además emplea ingeniería de modelos para establecer la metodología que permite parametrizar los niveles de estrés de los individuos. Se utilizó la librería Tweepy (Roesslein, 2015) de Phyton para la recolección de los Tweets que son analizados por el sistema.

El sistema presenta un reporte de valoración de los niveles del estrés para un determinado periodo de tiempo y región del Ecuador.

1.4. Objetivos

1.4.1. General

- Desarrollar una metodología basada en ingeniería de modelos que permita identificar los niveles de estrés de una persona aplicando Procesamiento de Lenguaje Natural.

1.4.2. Específicos

- Investigar y comparar técnicas de Procesamiento de Lenguaje Natural e ingeniería de modelos.
- Definir las herramientas de software de Procesamiento de Lenguaje Natural e ingeniería de modelos.
- Definir una metodología basada en modelos para la parametrización de los niveles de estrés.
- Desarrollar el algoritmo de Procesamiento de Lenguaje Natural para el análisis del corpus del texto.
- Diseñar e implementar un sistema capaz de detectar el nivel de estrés a partir de los mensajes escritos en la red social Twitter de la población ecuatoriana.
- Mostrar el reporte o histórico de los niveles de estrés en un determinado periodo de tiempo y región.

1.5. Estado del Arte

En el artículo “Desarrollo de un avatar animado con expresión de emociones básicas” (Medina, Starostenko, & Ruiz Castillo, 2013) se presenta un agente animado que tiene la capacidad de expresar emociones básicas a través de expresiones faciales con base en la teoría de Paul Ekman (Ekman & Friesen , 1978).

En el trabajo “Desarrollo del corpus de habla emocional tamil y evaluación mediante SVM” (Joe, 2014) se desarrolló un sistema de reconocimiento del habla los cuales tienen como objetivo reconocer la palabra hablada mediante la extracción de características que permiten reconocer las emociones en la señal de voz mediante el uso de corpus que son evaluados utilizando motores basados en SVM. El corpus del habla emocional se recoge en varios métodos como actuado, real y espontáneo. El corpus emocional debe ser analizado por más oyentes para comprobar la anotación de cada segmento como igual. Las bases de datos con diversas emociones se recogen y se analiza por las emociones y se etiquetan de acuerdo. Las emociones se categorizan en varios grupos emocionales, a saber: Feliz, Triste, Cólera, Miedo, Neural, etc. Para evitar ambigüedades entre las categorías, la base de datos es analizada por dos personas más en dos sesiones. Se descarta cualquier segmento que carezca de claridad en el contenido del lenguaje o en la intensidad de la señal. Las clasificaciones tienen que ser evaluadas por muchos oyentes. La evaluación la realizan dos tipos de oyentes, a saber, los oyentes acústicos y lingüísticos. Los oyentes acústicos observan el sonido por el que perciben.

En el proyecto “*Machine Learning Framework for the Detection of Mental Stress at Multiple Leve*ls” (Rauf Subhani, Mumtaz, Naufal, Kamel, & Saeed Malik, 2017) se informa sobre el desarrollo de instrumentación de hardware y software y el procesamiento de señal utilizado para detectar cambios en el nivel de estrés de un sujeto que interactúa con un ordenador, en el marco de una tarea experimental específica. Para este experimento se implementó un conjunto de computadora basado en un test de estrés mental clínico, llamado "*Stroop Test*", adaptado para hacer que el sujeto experimentara dos niveles diferentes de estrés, mientras que sus señales BVP, GSR y PD fueron registradas continuamente. Se aplicaron varias técnicas de procesamiento de datos para extraer atributos efectivos del estado de "estrés" de los sujetos.

CAPÍTULO 2

2.1. Introducción

El procesamiento del lenguaje natural o *NLP (Natural Language Processing*) nace como una rama de la Inteligencia Artificial y se ocupa de la capacidad de comunicación de los ordenadores con los humanos utilizando su propio lenguaje. Es un área cuyas aplicaciones son múltiples y variadas, como la traducción automática o el reconocimiento y comprensión del lenguaje humano entre otros (Gelbukh, 2010).

La lingüística computacional (o lingüística informática) es un campo científico interdisciplinar relativamente reciente cerca de cincuenta años de investigación y desarrollo cuyo objetivo radica en incorporar en los ordenadores la habilidad en el manejo del lenguaje humano. Desde el punto de vista de su vinculación a la informática, y también por motivos históricos, la lingüística computacional suele ser considerada como una subdisciplina de la inteligencia artificial. La inteligencia artificial, por su parte, es una subdisciplina de la informática que se ocupa de la comprensión de la inteligencia y del diseño de máquinas inteligentes, es decir, de máquinas que presentan características asociadas con el entendimiento humano, como el raciocinio, la comprensión del lenguaje hablado y escrito, el aprendizaje o la toma de decisiones. En una afortunada definición de (Minsky, 1967), la inteligencia artificial es "la ciencia de hacer que las máquinas hagan cosas que, de haber sido hechas por seres humanos, requerirían inteligencia" (Guinovart, 1998) (p. 135-146).

El procesamiento de lenguaje natural es un área de la inteligencia artificial y estudia lenguaje entre el humano y las maquinas informáticas permitiendo tener una comunicación de manera autónoma, estos prototipos abarcan muchas características del lenguaje humano. Para poder comprender el lenguaje humano se deben realizar ciertas acciones como son análisis morfológico, sintáctico, semántico y pragmático (Toledo Costa, Godoy Guerra, & Suárez Puente, 2008).

- **Análisis morfológico:** Consiste en analizar las palabras para obtener raíces, características flexivas, unidades léxicas compuestas y otros fenómenos.
- **Análisis sintáctico:** Consiste en analizar la estructura sintáctica de una oración mediante una gramática del lenguaje.

- **Análisis semántico:** Consiste en determinar el significado de una oración, y la resolución de ambigüedades léxicas y estructurales.
- **Análisis pragmático:** Consiste en analizar el texto más allá de los límites de la oración, por ejemplo, para establecer los antecedentes referenciales de los pronombres.

El concepto de estrés fue introducido por primera vez en el ámbito de la salud por Hans Selye (Selye, 1964), quién precisó el síndrome del estrés, como la réplica no específica derivada por provocaciones contrarios enormes y que denominó "Síndrome de estrés". Esta contestación está compuesta por tres etapas: una de miedo, otra de conciliación y otra de debilidad. En esta última causa un desgaste de los componentes adaptativos acompañado de señales de angustia (Sandín, 2003).

2.2. Procesamiento de Lenguaje Natural

El NLP se usa en varios campos de la ciencia y para varios fines como son el análisis lingüístico, traducción automática de textos, boots conversacionales, recuperación de información, análisis de emociones, clasificadores de texto.

2.2.1. Arquitectura básica de los sistemas de NLP

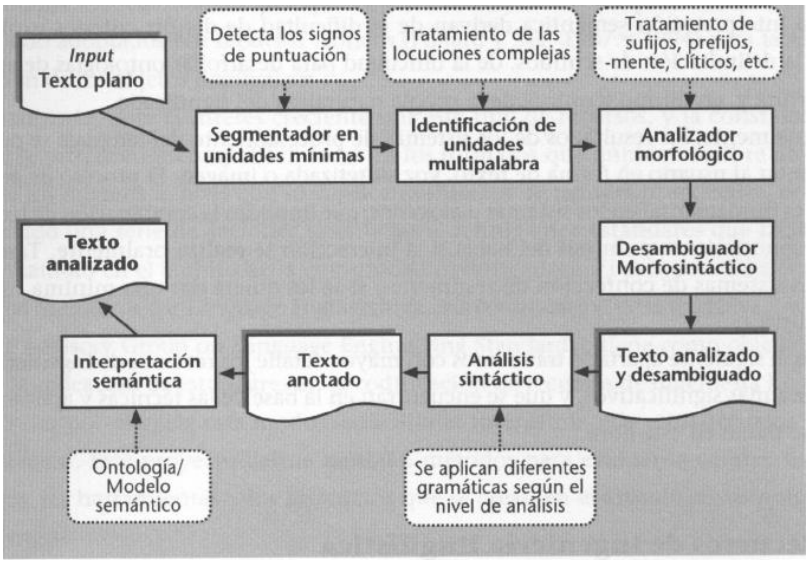

Figura 1. Arquitectura básica de un sistema de NLP
Fuente: (Martí, 2003)

2.2.2. Algoritmos de NLP orientados a la recuperación de información

La particularidad del rescate de la información radica en la exploración de documentos relevantes de una compilación dada una consulta que dice la escasez de información del usuario. Los documentos restituidos por un procedimiento de rescate de información son en general, regulados por grado de semejanza o relevancia relacionado a la consulta. Dicha labor de rescate se ejecuta, por defecto, por varios métodos:

• Estudio y regulación: escogimiento de los procesos que mejor simbolizan el contenido de los documentos y transformación de las técnicas elegidas con el propósito de comprimirlos a formas apropiadas que proporcionen los mensajes rezagados en la causa de búsqueda. Los requisitos pueden ser léxicos, expresiones, n-gramas, u otros mecanismos.

• Sistematización de pesos: concesión a cada uno de las cláusulas de un coste numérico que simboliza su jerarquía a la hora de simbolizar el contenido de un documento escrito.

• Indexación: creación de un índice que provea el camino a los documentos que contengan los requisitos que los simbolizan.

• Búsqueda: asunto basado en el cómputo de correspondencias y similitudes entre la representación de la consulta y la de cada documento. Para conseguir representaciones concurrentes y así aprobar contrastar consultas con documentos, el texto de cada sugestión deberá ser examinado de la misma forma que el de los documentos.

Para delinear un método de RI, existen muchos medios y disparidades en cuanto al tipo de rescate utilizado (Vilares, 2006). Los modelos más habituales son los siguientes:

• Tipo booleano, fundado en la hipótesis de conjuntos,

• Tipo vectorial, fundado en el álgebra (Salton & Buckley, 1988),

• Tipo probabilísticos, el más antiguo de arquetipo bayesiano (Robertson & Jones, 1976)frente a los fundados en modelos del lenguaje (Ponte & Croft, 1998).

Una opción sugestiva para optimizar el proceso de rescate se centra en la diversión de las consultas. Diversas tácticas han sido expuestas:

• Realimentación por preeminencia, en el que la consulta originaria se expande con términos extraídos de los documentos más notables regenerados a partir de la sugestión inicial.

• Esparcimiento con texto completo, donde el esparcimiento se realiza con textos exactos que contienen las palabras de la consulta originaria, acopiados tanto de textos notables como no distinguidos (Strzalkowski, Lin, Perez Carballo, & Wang, 1997).

• Esparcimiento con sinónimos, donde los términos de la consulta inicial se amplían con sinónimos ordinariamente obtenidos de fuentes léxicas como WordNet (Miller, 1995). De forma opuesta de los dos métodos preliminares, que han dado en forma frecuente excelentes resultados, la expansión con equivalentes no ha ayudado a optimizar los métodos de recuperación de información.

2.2.3. Términos usados en NLP

Los términos más comunes utilizados en NLP son los siguientes:

- **Token:** Este término representa a una palabra del texto y es la unidad más simple de procesamiento.
- **Tokenización:** Este proceso consiste en dividir una sentencia en cada uno de los tokens que la componen. Aunque puede ser un proceso simple para textos escritos en algunas lenguas, especialmente en el caso de las románicas cuyo token separador es un espacio, en otras lenguas como las altaicas, extraer los tokens de una sentencia es un proceso mucho más complejo debido a la sintaxis y semántica de los escritos en dichas lenguas.
- **Sentencia:** Este término representa una oración conformada por varios tokens.
- **Corpus:** Este término representa al cuerpo del mensaje que se encuentra compuesto por un conjunto de sentencias.
- **Part-of-speech (POS):** Dependiendo de la semántica del lenguaje, cada token que compone una sentencia puede ser un verbo, un adjetivo, un pronombre, un artículo, etc. Un POS es simplemente una clasificación para cada token dentro de una sentencia, de esta forma es posible identificar el significado de cada token y las partes clave de cada sentencia.
- **Stemming:** Este proceso consiste en la eliminación de los prefijos y sufijos de una palabra para obtener la palabra raíz de la cual se origina.

2.3. Minería de opinión

Se puede precisar a la minería de datos o estudio de emociones a la aplicación del proceso de lenguaje natural para asemejar y extraer la información intrínseca de un contenido.

La idea detrás del estudio de emociones es establecer la cualidad del autor de una obra respecto a un explícito tema. Esta cualidad puede ser su propia apreciación, estado afable (estado sentimental cuando creo la obra) o la fase sentimental que desea imprimir al leyente o consumidor de tal obra.

Los trabajos en esta área están encaminados a establecer la polaridad de un texto intrínseco, esto es, resolver si un texto es positivo o negativo usando para ello diferentes métodos, desde la detección de adjetivos que conlleven una dimensión sentimental a la categorización automática por medio el uso de textos previamente inscritos.

La detección de polaridad mediante calificaciones se remonta a 1971, cuando McNair, Lorr y Droppleman establecieron la lista POMS (Profile of Mood States), consistente en 65 calificaciones predestinados a calcular siete elementos de los estados de ánimo: tensión, depresión, ira, vigor, fatiga, desconcierto y afecto. Inicialmente esbozada para el ámbito clínico, su uso se ha desarrollado a otros campos como la lingüística computacional. También existe un ajuste de dicha lista de adjetivos al español, no indulta de dificultades semánticos durante su transcripción (Fernández & Pesqueira, 2000).

En cuanto al uso de textos inscritos, Peter Turney y Bo Pang fueron los primeros escritores en ocuparse en esta área usando reseñas de bienes (Turney, 2002) y de películas (Pang, Lee, & Vaithyanathan, 2002) correspondientemente.

Otro de los objetivos de la minería de datos es la identificación de textos objetivos y subjetivos, una labor si cabe más complicada que la determinación de la polaridad ya que la subjetividad de las palabras o frases dependen del contexto en el que se hallen siendo muy usual que textos objetivos contengan información intrínseca.

2.3.1. Twitter

Consiste en un servicio de red social americano fundado en 2006 con el propósito de facilitar el intercambio de mensajes de texto breves entre sus miembros.

Considerado por numerosas personas como un servicio de micro blogging, su vital virtud junto a la simplicidad de uso y simplicidad, es la limitación de los mensajes que toman nombre de tweet a tan solo 140 caracteres como máximo, lo que ha supuesto que algunos medios y autores describan este instrumento como los "SMS de la web" (D'Monte, 2017).

Esta situación en la longitud de los mensajes lo hace ideal para intercambiar contenido de forma rápida, el cual se acumula de forma constante en los servidores de Twitter, y tiene su definición en el origen de este instrumento.

Twitter se comenzó a fraguar en 2006 en el seno de una compañía dedicada a los podcasts (archivos multimedia por suscripción) como idea para comunicar a pequeños grupos de personas partiendo del conocimiento de los *SMS (Short Message Service)* en telefonía móvil (Sarno, 2009). A esta idea la llamaron "*twttr"* y tras un tiempo de experimentos internos, fue lanzada al público el 15 Julio de 2006 (Arrington, 2006).

Mas tarde, los creadores de la idea se reorganizaron en una nueva empresa, creando *Obvious Corporation* y absorbiendo a la antigua compañía de podcast para más adelante refundar la compañía como Twitter Inc. La red social se hizo especialmente popular en 2007, en el evento SXSW (South by Southwest Interactive) en el que se proyectaron los mensajes de los usuarios en varios televisores repartidos por las instalaciones lo cual ánimo a los asistentes al congreso a participar, consiguiendo triplicar en número de tweets en apenas unos días (Shi, Rui, & Whinston, 2013).

Al inverso de lo que sucede en otras redes sociales, en Twitter no es obligatorio que exista una correlación recíproca entre sus miembros para compartir información, sino que ésta puede ser desigual, no requiriendo un consentimiento mutuo entre ambos.

Este comportamiento se iguala mucho más al mundo real, en el que la comunicación entre objetos puede ser unidireccional como ya ocurre con los medios de comunicación tradicionales: prensa, radio y televisión.

Las relaciones entre los usuarios de Twitter pueden ser de dos tipos: *following* y follower, que podría traducirse como seguidor y seguido. En efecto, se llama *following*

al conjunto de miembros a los que otro usuario sigue, suscribiéndose a sus publicaciones.

Del modo equivalente, estos usuarios seguidos pasan ahora a disponer de un nuevo seguidor o *follower*. Si dos usuarios se siguen recíprocamente, se podría pensar entonces que hay una relación mutua de "afecto", similar a la que se da en otras redes sociales.

2.3.2. Test psicológico de Sacks

El Test de Frases Incompletas (Sacks & Levy, 1950) es parecido al método de asociación de palabras con ciertas variaciones. Estas técnicas han sido evaluadas y a menudo, se han obtenido mejores resultados con el test de Sacks. Estos resultados muestran que el test reduce la diversidad de las asociaciones invocadas por una palabra, propone de mejor manera los contextos, tonos, cualidades de la actitud y las áreas de atención, permitiendo mayor libertad y mayores opciones de respuesta, abarca campos más específicos y definidos de la actitud de la persona.

En este test se espera que la persona muestre sus propios deseos, sentimientos, temores y actitudes en las frases que forma. Se ha encontrado que con el test de Sacks se revelan pensamientos conscientes, inconscientes y preconscientes de la persona, que el psicólogo puede analizar y comparar con resultados obtenidos por medio de otras técnicas.

A diferencia de los test objetivos o estructurados, el test de Sacks presenta la gran ventaja de dar una libertad de respuesta, en vez de limitarla a contestar de manera afirmativa con un "SI" o negativa con un "NO", la persona puede responder al estímulo como desee.

De esta forma, la naturaleza del test queda encubierto, ya que la persona no sabe a ciencia cierta qué respuesta es "buena" o cual "mala", aunque intuya o incluso conozca con qué propósito se le aplica el test. El test se puede interpretar en forma cualitativa y cuantitativa. El test de Sacks consta de 60 frases incompletas que se enfocan en cuatro diferentes áreas de estudio área de adaptación familiar, área sexual, área de relaciones interpersonales, y área de autoconcepto:

2.3.2.1. Área de adaptación familiar

Tabla 2
Frases del área de adaptación familiar

Actitud hacia el padre (Frases 1, 16, 31 y 46)	
Número	**Frase incompleta**
1	Siento que mi padre raras veces
16	Si mi padre tan sólo
31	Desearía que mi padre
46	Siento que mi padre es
Actitud hacia la madre (Frases 14, 29, 44 y 59)	
Número	**Frase incompleta**
14	Mi madre
29	Mi madre y yo
44	Creo que la mayoría de las madres
59	Me agrada mi madre, pero
Actitud hacia la unidad familiar (Frases 12, 27, 42 y 57)	
Número	**Frase incompleta**
12	Comparada con las demás familias, la mía
27	Mi familia me trata como
42	La mayoría de las familias que conozco
57	Cuando era niño (a), mi familia

Fuente: (Sacks & Levy, 1950)

A través de las 12 frases que componen esta área, la persona va a expresar sus sentimientos hacia cada uno de los padres por separado, y hacia la familia como un todo.

2.3.2.2. Área sexual

Tabla 3
Frases del área sexual

Actitud hacia los hombres/las mujeres (Frases 10, 25, 40 y 55)	
Número	**Frase incompleta**
10	Mi idea de mujer (hombre) perfecta (o)
25	Pienso que la mayoría de las muchachas (os)
40	Creo que la mayoría de las mujeres (hombres)
55	Lo que menos me gusta de las mujeres (hombres)
Actitud hacia las relaciones heterosexuales (Frases 11, 26, 41 y 56)	
Número	**Frase incompleta**
11	Cuando veo un hombre y a una mujer juntos
26	Yo creo que la vida matrimonial
41	Si tuviera relaciones sexuales
56	Mi vida sexual

Fuente: (Sacks & Levy, 1950)

A través de las 8 frases que componen esta área, el examinado expresará su actitud hacia el sexo opuesto, hacia el matrimonio y las relaciones sexuales.

2.3.2.3. Área de relaciones interpersonales

Tabla 4
Frases del área de relaciones interpersonales

Actitud hacia amigos y conocidos (Frases 8, 23, 38 y 53)	
Número	**Frase incompleta**
8	Creo que es un verdadero amigo
23	No me gusta la gente con
38	La gente que más me agrada
53	Cuando no estoy, mis amigos
Actitud hacia colegas en el trabajo o escuela (Frases 13, 28, 43 y 58)	
Número	**Frase incompleta**
13	En las labores me llevo mejor con
28	Aquellos con los que trabajo

Continua ⟶

43	Me gusta trabajar con la gente que
58	La gente que trabaja conmigo, generalmente
Actitud hacia superiores en el trabajo o escuela (Frases 6, 21, 36 y 51)	
Número	**Frase incompleta**
6	Las personas que están sobre mí
21	En la escuela, mis maestros
36	Cuando veo el jefe venir
51	La gente a quien yo considero mis superiores
Actitud hacia los subordinados (Frases 4, 19, 34 y 49)	
Número	**Frase incompleta**
4	Si yo estuviera a cargo
19	Si la gente trabaja para mí
34	La gente que trabaja para mí
49	Lo que más deseo en la vida

Fuente: (Sacks & Levy, 1950)

A través de las 16 frases que componen esta área, el examinado expresará sus sentimientos hacia personas fuera de su hogar, y su idea de lo que sienten los demás con respecto a él.

2.3.2.4. Área de autoconcepto

Tabla 5
Frases del área de autoconcepto

Actitud hacia los temores (Frases 7, 22, 37 y 52)	
Número	**Frase incompleta**
7	Sé que es tonto, pero tengo miedo de
22	La mayoría de mis amistades no saben que tengo miedo de
37	Quisiera perder el miedo de
52	Mis temores en ocasiones me obligan a

Continua ⟶

Actitud hacia los sentimientos de culpa (Frases 15, 30, 45 y 60)	
Número	**Frase incompleta**
15	Haría cualquier cosa por olvidar la vez que
30	Mi más grande error fue
45	Cuando era más joven me sentía culpable de
60	La peor cosa que he hecho
Actitud hacia las metas (Frases 3, 18, 33 y 48)	
Número	**Frase incompleta**
3	Siempre anhelé
18	Sería perfectamente feliz si
33	Mi ambición secreta en la vida
48	Cuando doy órdenes, yo
Actitud hacia las propias capacidades (Frases 2, 17, 32 y 47)	
Número	**Frase incompleta**
2	Cuando tengo mala suerte
17	Siento que tengo la habilidad para
32	Mi mayor debilidad es
47	Cuando la suerte se vuelve en contra mía
Actitud hacia el pasado (Frases 9, 24, 39 y 54)	
Número	**Frase incompleta**
9	Cuando era niño (a)
24	Antes
39	Si fuera joven otra vez
54	Mi más vivido recuerdo de la infancia

Continua ⟶

Actitud hacia el futuro (Frases 5, 20, 35 y 50)	
Número	**Frase incompleta**
5	El futuro me parece
20	Yo espero
35	Algún día yo
50	Dentro de algún tiempo

Fuente: (Sacks & Levy, 1950)

A través de las 24 frases que componen esta área, la persona dará a conocer un cuadro del concepto que tiene de sí mismo tal cómo es, cómo fue, cómo espera ser y cómo cree realmente que será.

2.4. Síndrome del estrés

En 1936 Selye precisó el significado de estrés ante la OMS, como la reacción no determinada del cuerpo ante cualquier exigencia o demanda del exterior. Es decir, una respuesta global a escenarios externos que turban el equilibrio sentimental y funcional del individuo (Slipak, 1991).

Un axioma más, señala que el estrés es el conjunto de métodos y respuestas neuroendocrinas, inmunológicas, sentimentales y de conducta, ante circunstancias que significan una demanda de adaptación mayor que lo habitual para el cuerpo y que son descubiertas por el individuo como una advertencia o riesgo para su integridad biológica o psíquica (Cassaretto, Chau, Oblitas, & Valdez, 2003).

El estrés consigue afectar a cualquier persona sin distinción de raza, sexo, edad o posición social. Las primeras manifestaciones suelen ser depresión, dolores musculares, disturbios del sueño, pérdida del apetito, pérdida del interés y de la concentración. Estudios epidémicos y sociales han confirmado que niveles altos de estrés guardan relación con las enfermedades y la mortalidad de la población, ya sea por perturbaciones cerebrales, incidentes, intimidación, cáncer, contagios o padecimientos cardiovasculares. El estrés en un elemento de peligro para padecimiento de coronaria y malestar cerebro vascular, patologías de elevada prevalencia en las mujeres en etapa de menopausia y sobre todo en la posmenopausia (Villamil Gómez, y otros, 2015).

El síndrome del estrés o Síndrome de Adaptación General de Selye también se define como la respuesta del organismo ante una situación de estrés ambiental distribuida en tres fases o etapas la fase de alarma, la fase de resistencia y la fase de agotamiento (Duval, González, & Rabia, 2010).

2.4.1. Etapa de alarma.

Ante la percepción de una potencial situación de estrés, el cuerpo empieza a desarrollar una serie de modificaciones de orden orgánico y psíquico (ansiedad, inquietud, etc.) que lo inducen a enfrentar la situación estresante. La aparición de estos indicios está influida por componentes como las cuantificaciones físicas del estímulo ambiental, elementos de la persona, el nivel de amenaza percibido y distintos como el grado de control sobre el estímulo o la presencia de otros estímulos ambientales que influyen sobre el escenario.

2.4.2. Etapa de resistencia.

Supone la fase de adaptación a la situación estresante. En ella se desarrollan un conjunto de procesos fisiológicos, cognitivos, emocionales y comportamentales destinados a "negociar" la situación de estrés de la manera menos lesiva para la persona. Si finalmente se produce una adaptación, esta no está exenta de costos, p.e. disminución de la resistencia general del organismo, disminución del rendimiento de la persona, menor tolerancia a la frustración o presencia de trastornos fisiológicos más o menos permanentes y también de carácter psicosomático.

2.4.3. Etapa de agotamiento.

Si la fase de resistencia fracasa, es decir, si los mecanismos de adaptación ambiental no resultan eficientes se entra en la fase de agotamiento donde los trastornos fisiológicos, psicológicos o psicosociales tienden a ser crónicos o irreversibles.

2.5 Clasificadores de texto

El objetivo de los clasificadores de texto es hallar propiedades y rasgos del lenguaje para establecer un tipo o categoría a un texto escrito.

La forma más común es la categorización de textos por tópico a través de la extracción de información objetiva.

2.5.1. Clasificadores de texto tipo Ad-Hoc

Uno de las formas más simples es estableciendo un conjunto de normas según las cuales se determina la categoría o tipo de texto. Un caso particular es el de análisis de sentimientos en donde se tienen un grupo de palabras, sentencias o características ad-hoc que sugieran puntos de vista positivos o negativos y en base a ello analizar y categorizar el texto (Angeli, Manning, & Jurafsky, 2012).

2.5.2 Clasificadores basados en Machine Learning

Uno de los campos de estudio que cada vez está obteniendo más popularidad dentro de las ciencias de la informática es el aprendizaje automático o Machine Learning. El Machine Learning, o Aprendizaje Automático en español se define como la capacidad que tiene una computadora para aprender a pensar como un humano (Samuel, 1969).

2.5.3. Clasificador Supervisado

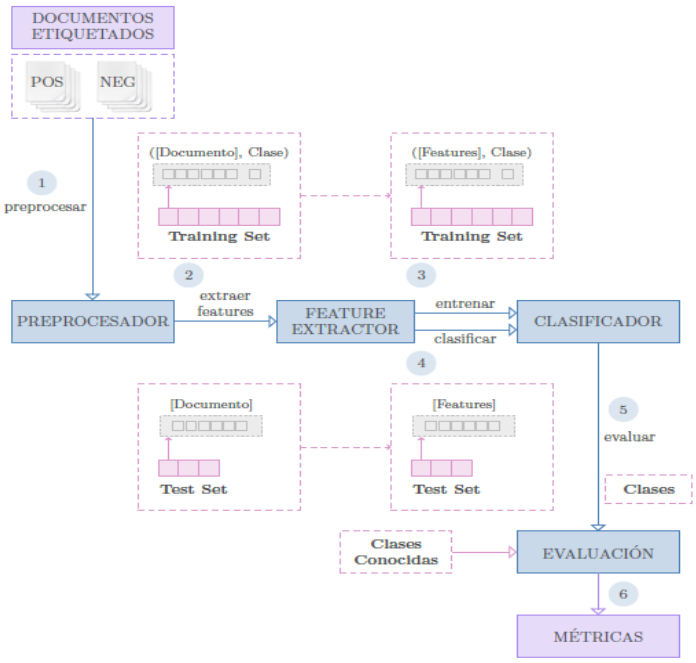

Figura 2 Proceso de Clasificación Supervisado
Fuente: (Samuel, 1969)

En los métodos supervisados la clasificación depende del entrenamiento que se realice sobre el clasificador por lo que su efectividad dependerá en gran medida de que datos del corpus se utilicen para entrenamiento (training set) y cuales para evaluarlo (test set). El criterio de selección de estos conjuntos de datos y la cantidad de veces que se evalué al clasificador utilizando distintos datos definirá la confiabilidad de la evaluación (Martínez Cámara, Valdivia, Teresa, Perea Ortega, & Ureña López, 2011).

2.5.3.1. Clasificador Naive Bayes

Naive Bayes es un método de clasificación supervisado y generativo que se basa en el teorema de Bayes y en la premisa de independencia de los atributos dada una clase. Esta premisa es conocida como "Naive Assumption" y se le llama "naive" (o ingenua) considerando que en la práctica los atributos raramente son independientes, lo cual, en la mayoría de los casos, no afecta los resultados del método.

Considerando el teorema de Bayes,

$$\mathrm{P(C_i|D)} = \frac{\mathrm{P(C_i)P(D|C_i)}}{\mathrm{P(D)}} \qquad (1)$$

Donde *D* es un documento del conjunto de datos de entrenamiento, *Ci* es cada una de las posibles clases y $P(C_i|D)$ es la probabilidad de que el documento *D* pertenezca a la clase Ci. El clasificador seleccionará la clase que maximice esta probabilidad.

P(D) puede ser ignorada ya que es la misma para todas las clases y no afecta los valores relativos de probabilidad.

Además, basados en la premisa de independencia de los atributos dada una clase podemos descomponer $P(D|C_i)$ como se ve en la ecuación que sigue:

$$\mathrm{P(C_i|D)} \propto \mathrm{P(C_i)} \prod_{\mathrm{k=1}}^{\mathrm{n}} \mathrm{P(f_k|C_i)} \qquad (2)$$

Donde f_k son los features del documento y $P(f_k|C_i)$ es la probabilidad de ocurrencia del feature en la clase dada.

Por lo tanto, la clase seleccionada por el clasificador será la que maximice la probabilidad anterior.

$$C_{NB} = \arg max_i \, P(C_i) \prod_{k=1}^{n} P(f_k|C_i) \qquad (3)$$

Las distintas implementaciones del algoritmo de Naive Bayes difieren principalmente en la aproximación de $P(f_k|C_i)$ y las técnicas de smoothing utilizadas para el tratamiento de probabilidades bajas o nulas (Martínez Cámara, Valdivia, Teresa, Perea Ortega, & Ureña López, 2011).

2.5.3.2. Multinomial Naive Bayes

Consideremos el caso de utilizar como features todos los términos del vocabulario del corpus de entrenamiento teniendo en cuenta su frecuencia de aparición; add-one como técnica de smoothing; y MLE para estimar $P(f_k|C_i)$. Este caso es conocido como Naive Bayes multinomial y la implementación será como sigue:

$$\widehat{\mathrm{P}}(\mathrm{w_i}|\mathrm{C_i}) = \frac{\mathrm{count}(\mathrm{w_i},\mathrm{C_i})+1}{|\mathrm{V}|+\sum_{\mathrm{w}}\mathrm{count}(\mathrm{w},\mathrm{C_i})} \qquad (4)$$

Donde, $count(w, C_i)$ será la cantidad de veces que el término w_i aparece en la clase C_i y $\sum_w count(w, C_i)$ será la sumatoria de frecuencias de aparición de cada término en la clase Ci.

Una consideración importante a tener en cuenta en la implementación de algoritmos que calculan probabilidades es que conviene realizar los cálculos en forma logarítmica porque las probabilidades suelen ser valores muy pequeños y al multiplicarlos podemos encontrarnos con un problema de precisión. Por otro lado, esta forma de cálculo tiene la ventaja de que computacionalmente la suma es menos costosa que la multiplicación (Kibriya, Frank, Pfahringer, & Holmes, 2004).

2.5.4. Clasificador No Supervisado

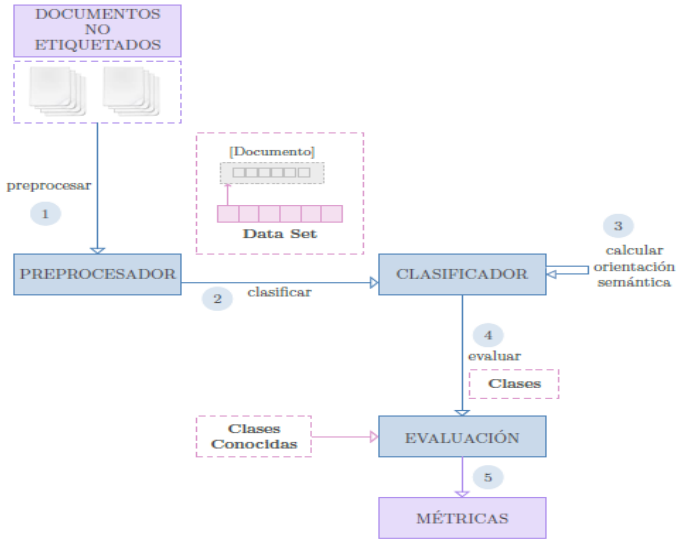

Figura 3 Proceso de Clasificación No Supervisado
Fuente: (Samuel, 1969)

En algoritmos de aprendizaje no supervisado el entrenamiento se realiza a partir de una gran colección de experiencias y atributos de esas experiencias, pero no se tienen datos sobre los resultados correctos, es decir, no se conoce la clase a la que pertenece cada experiencia. En este tipo de algoritmos el objetivo es encontrar similitudes, patrones o estructuras en los datos que permitan agruparlos sin conocer previamente los grupos existentes.

2.5.5. Métricas de Evaluación de Clasificadores

Dependiendo de la tarea de procesamiento de lenguaje que se esté desarrollando podría variar el significado o el sentido que se le da a cada medición, pero siempre se consideraran los siguientes casos (Laza & Pavón, 2010):

- **Verdaderos Positivos**, llamados (VP): Cuando un texto es clasificado como positivo y este era efectivamente positivo.
- **Falsos Positivos**, llamados (FP): Cuando un texto es clasificado como positivo, pero este era negativo.
- **Falsos Negativos,** llamados (FN): Cuando un texto es clasificado como negativo, pero este era positivo.
- **Verdaderos Negativos**, llamados (VN): Cuando un texto es clasificado como negativo y este era efectivamente negativo.

Tabla 6
Evaluación del clasificador

Clasificación	Positivo	Negativo
Texto clasificado como positivos	VP	FP
Texto clasificado como negativo	FN	VN

Fuente: (El Autor, 2017)

2.5.5.1. Cálculo de la efectividad en clasificadores de texto

A continuación, se muestran las métricas para determinar la efectivad de un clasificador (Forman, 2003):

- **Accuracy** Es la cantidad total de textos que son clasificados como positivos y como negativos de forma correcta sobre el total de textos.

$$Accuracy = \frac{VP+VN}{VP+VN+FP+FN} \quad (5)$$

- **Precisión de clasificación de textos positivos** Es la cantidad de textos que son clasificados como positivos de forma correcta sobre el total de textos positivos.

$$PresiciónPOS = \frac{VP}{VP+FP} \quad (6)$$

- **Precisión de clasificación de textos negativos** Es la cantidad de textos que son clasificados como negativos de forma correcta sobre el total de textos negativos.

$$PresiciónNEG = \frac{VN}{VN+FN} \quad (7)$$

- **Recall de clasificación de textos positivos** Es la cantidad de textos de la categoría de positivos que son relevantes para el clasificador.

$$RecallPOS = \frac{VP}{VP+FN} \quad (8)$$

- **Recall de clasificación de textos positivos** Es la cantidad de textos de la categoría de negativos que son relevantes para el clasificador.

$$RecallNEG = \frac{VN}{VN+FP} \quad (9)$$

Para encontrar una métrica que represente la efectividad de un clasificador, teniendo en cuenta tanto la precisión como la recall, de modo que nos permita tener una idea más general de cuan buenos son los resultados, se utilizó lo que se conoce como F-Measure (Powers, 2011).

F-Measure combina las medidas de precisión y recall a partir de la media armónica ponderada de estos dos valores.

$$F = \frac{1}{\propto\frac{1}{Precisión}+(1-\propto)\frac{1}{Recall}} = \frac{(\beta^2+1)Precisión*Recall}{\beta^2 Precisión+Recall} \quad (10)$$

Por lo general se utiliza una forma balanceada de la métrica anterior, es decir, para $\beta = 1$ conocida como F1-Measure.

$$F1_{POS} = \frac{2*Precisión_{POS}*Recall_{POS}}{Precisión_{POS}+Recall_{POS}} \quad (11)$$

$$F1_{NEG} = \frac{2*Precisión_{NEG}*Recall_{NEG}}{Precisión_{NEG}+Recall_{NEG}} \quad (12)$$

2.5.5.2. Ejemplo de medición de la efectividad en un clasificador de texto

Tabla 7
Ejemplo de clasificación de textos

Texto	Categoría asignada manualmente	Clasificación del algoritmo	Caso
Texto 1	NEG	NEG	VN
Texto 2	NEG	NEG	VN
Texto 3	NEG	NEG	VN
Texto 4	NEG	NEG	VN
Texto 5	NEG	NEG	VN
Texto 6	POS	NEG	FN
Texto 7	POS	POS	VP
Texto 8	NEG	NEG	VN

Continua ⟶

Texto 9	NEG	POS	FP
Texto 10	NEG	NEG	VN
Texto 11	POS	NEG	FN
Texto 12	POS	NEG	FN
Texto 13	POS	NEG	FN
Texto 14	POS	NEG	FN
Texto 15	POS	POS	VP
Texto 16	NEG	NEG	VN
Texto 17	NEG	NEG	VN
Texto 18	NEG	NEG	VN
Texto 19	NEG	NEG	VN
Texto 20	NEG	NEG	VN

Fuente: (El Autor, 2017)

Tabla 8.
Sumatorio total de casos

Clasificación	**Positivo**	**Negativo**
Textos clasificados como positivos	VP=2	FP=1
Textos clasificados como negativos	FN=5	VN=12

Fuente: (El Autor, 2017)

Remplazando los valores de VP, FP, FN, y VN en las ecuaciones (5), (6), (7), (8) y (9), se obtienen los siguientes resultados.

AccuracyTOTAL=0.7

PrecisiónPOS =0.66

RecallPOS=0.286

PrecisiónNEG=0.706

RecallNEG=0.923

Esto quiere decir que:

- El 70% del total de textos fueron clasificados correctamente en la categoría que pertenecen(AccuracyTOTAL).

- Del total de textos positivos solamente el 66% fueron clasificados correctamente (PrecisiónPOS).
- Del total de textos positivos, solamente el 28% son relevantes para el clasificador (RecallPOS).
- Del total de textos negativos solamente el 70% fueron clasificados correctamente (PrecisiónNEG).
- Del total de textos negativos, el 92% son relevantes para el clasificador (RecallNEG).

Remplazando los valores de Precisión y Recall del ejemplo anterior en las ecuaciones (11) y (12).

$$F1_{POS} = \frac{2 * 0.66 * 0.286}{0.66 + 0.286} = 39.9\%$$

$$F1_{NEG} = \frac{2 * 0.706 * 0.923}{0.706 + 0.923} = 80\%$$

De acuerdo a los valores de F-Measure obtenidos se concluye que la efectividad de clasificación de textos negativos es mejor que la de textos positivos para este ejemplo.

2.6. Tecnologías usadas para el desarrollo del sistema

2.6.1. Python

Python es un lenguaje de programación interpretado de alto nivel, multipropósito y multiparadigma cuya filosofía de diseño enfatiza la legibilidad del código, contando con una sintaxis clara y expresiva, que junto a su extensa librería estándar hace que programar sea más rápido y productivo que en otros lenguajes como C, C++ o Java (Perone, 2009).

Entre sus características, destacan:

- Sintaxis clara y legible.
- Orientación a objetos muy intuitiva.
- Modulable, incluyendo soporte de paquetes jerárquicos.
- Gestión de errores basado en excepciones.
- Soporte de meta clases, decoradores y tipado dinámico de los datos.
- Permite al desarrollador escribir sus programas de manera potente y rápida.

- Potente y extensa librería estándar y gran cantidad de módulos de terceros para casi cualquier tarea.
- Posibilidad de emplear módulos escritos en diferentes lenguajes como C, C++ o Java.

Python soporta múltiples paradigmas de programación, como la programación orientada a objetos o la programación funcional e imperativa. Igualmente cuenta con una gestión de memoria automática similar a la de otros lenguajes interpretados como Ruby, Scheme, Perl o Tcl y se ofrece para múltiples sistemas operativos haciendo que un programa se comporte de la misma forma en varias plataformas siempre que cuenten con el intérprete apropiado.

Creado a finales de 1980 por Guido van Rossum en el Centro para las Matemáticas y la Informática, en los Países Bajos, quien puso nombre al lenguaje en honor a los satíricos anglos Monty Python. No fue hasta el lanzamiento de la versión 2.0 en octubre de 2000 cuando el lenguaje se concibió realmente popular concordando con una serie de permutas entre los que se incluía la apertura del desarrollo a la comunidad a través de los *Python Enhancement Proposals*, documentos para describir guías de estilo y nuevas peculiaridades propuestos por la colectividad, que son perennemente examinados hasta que se alcance un asentimiento.

Aunque el progreso del proyecto se ha llevado a cabo con la versión 2.6.5, disponible en el sistema operativo Ubuntu 10.04, actualmente se continua el desarrollo de Python en su versión 3.3 igualmente conocido como Python3000, a la cual recomiendan emigrar por su estabilidad y nuevas particularidades a pesar de no ser compatible con el software actual.

Tal es la versatilidad de Python, que existen cuantiosas implementaciones del lenguaje que podemos manejar según nuestras necesidades:

- CPython es la culminación original que se brinda en el mercado oficial de Python. Escrita en C.
- Jython, implementación escrita en Java que admite importar cualquier clase Java y compilar a bytecode de ese lenguaje.
- IronPython hace lo propio para .NET y Mono. Escrita en C#.

- PyPy, interprete y compilador JIT escrito en Python, más vertiginoso y eficaz que CPython.

Python está gobernado por la *PSF*, una organización sin ánimo de beneficio dedicada a la expansión de este lenguaje de programación, tramitando el progreso de nuevas versiones y la obtención de inversión. Python posee una licencia de código abierto, nombrada *Python Software Foundation License* que es compatible con la GNU a partir de la versión 2.1.1.

2.6.2. Tweepy

Se trata de un módulo escrito en Python para acceder e interactuar con la API de Twitter, desarrollado por Joshua Roesslein y liberado a la comunidad a través de GitHub mediante la licencia MIT (Roesslein, 2015).

Se podría decir que tweepy es un envoltorio de la API REST de Twitter que encapsula las peticiones HTTP como funciones simples en Python de forma que el desarrollador no se tiene que preocupar de lo que haya por debajo. Gracias a esta librería es posible utilizar las funcionalidades de Twitter en nuestra aplicación, capturando los tweets públicos que coincidan con los parámetros especificados en cada experimento, empleando para ello el llamado Streaming API.

Durante el desarrollo del proyecto, Twitter ha efectuado numerosos cambios en su API, desde la forma en la que se devolvían los resultados a la forma de autenticar las llamadas a la misma, pasando por un cambio de API general que tuvo lugar en junio de 2013, teniendo que actualizar consecuentemente el módulo para poder mantener la aplicación funcionando.

Así, la última versión de este módulo empleado en el despliegue del proyecto es la 2.0 con soporte para la API 1.1 de Twitter. Esta API es usada para extraer una gran cantidad de mensajes o tweets en tiempo real.

La API usa una instancia llamada tweepy Stream que establece una sesión de streaming o transmisión de datos para enviar los mensajes a otra instancia llamada StreamListener.

La función on_data() de un stream listener recibe todos los mensajes y llama a las funciones de acuerdo al tipo de mensaje. Para poder usar la API se realizan los siguientes pasos:

- Se crea una clase heredada de la clase StreamListener
- Se usa esta clase para crear un objeto tipo Stream
- Conexión a Twitter mediante la clase Stream

2.6.2.1. Creación de un StreamListener

El stream listener permite imprimir el estado del texto. El método on_status() muestra el estado de los datos. Se usa la clase MyStreamListener para la creación del flujo StreamListener.

2.6.2.2. Creación de un Stream

Para comenzar la causa es preciso registrar la aplicación de consumidor de Twitter, crear una nueva aplicación y conseguir los recetarios de consumidor a través los siguientes pasos:

- Conseguir una solicitud a partir de Twitter
- Redireccionar el consumidor a twitter.com para autorizar la aplicación.
- Permiso de acceso.

Una vez que hayamos logrado registrar el API se puede crear un objeto de tipo stream el cual permitirá ejecutar el enlace hacia la red social Twitter.

Uno de las actualizaciones más sonadas llevadas a cabo por Twitter, es el de la forma en la que se certificaban las llamadas a la API. Hasta junio de 2013, estos llamamientos podían certificarse o bien con el nombre de usuario y la contraseña de cada órgano de la red social transferida de forma segura por medio de HTTPS o bien mediante, un protocolo libre llamado OAuth que consiente una autorización segura para el uso de un API sin necesidad de meter la contraseña.

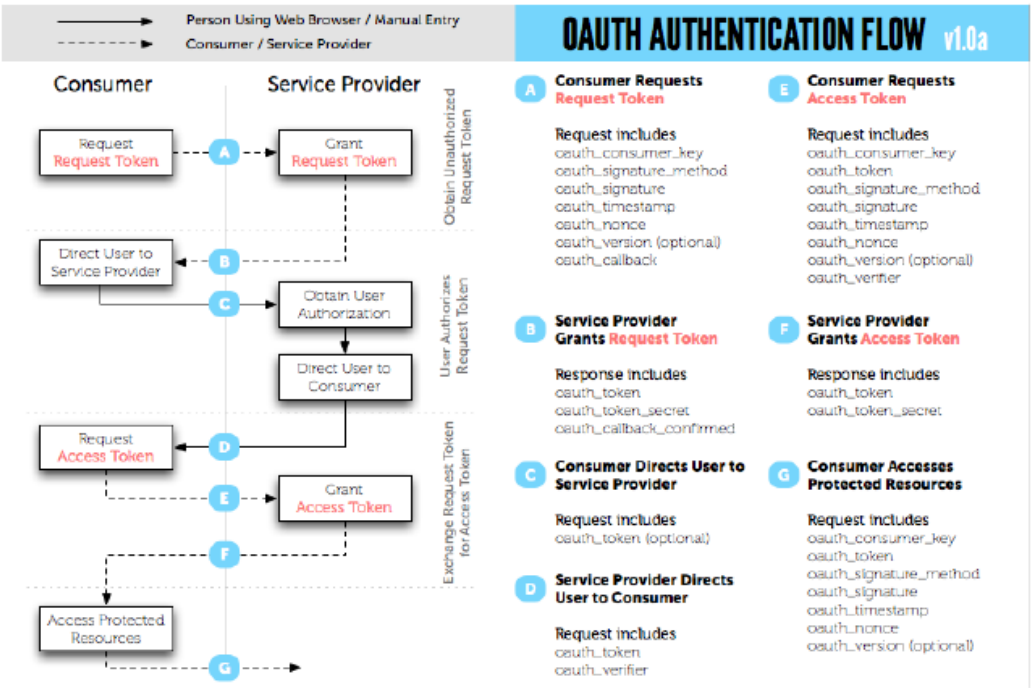

Figura 4. Diagrama de autorización de uso de la API de Twitter
Fuente: (Twitter Inc., 2016)

A partir de junio de 2013 solo se consiente interactuar con Twitter por medio de OAuth, cuyo soporte es prestado de forma íntegra por Tweepy, asumiendo que registrar todas las apis en el portal de desarrolladores de Twitter.

Con este método de permisión, los consumidores de Twitter deben conceder su aprobación expresa a que las aplicaciones aprueben a las funciones de su cuenta. Dicha aprobación es concedida o rechazada en la propia web de la red social, a la que serán redireccionados durante la interacción con cada una de las aplicaciones.

Cuando una aplicación es acreditada, Twitter le entrega las claves (access token key y access token secret) necesarias para efectuar acciones en la red social en nombre del usuario sin necesidad de solicitar su clave.

Esta metodología de permisión tiene una gran ventaja para los usuarios, ya que así evitan entregar la clave de sus cuentas a las aplicaciones, consiguiendo además revocar el acceso concedido anticipadamente a cada aplicación en cualquier instante.

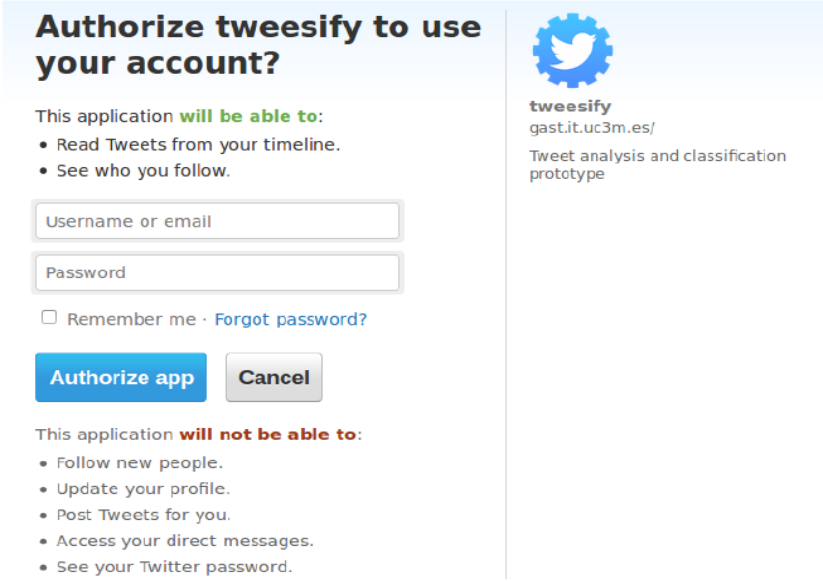

Figura 5. Página de autorización OAuth en Twitter
Fuente: (Twitter Inc., 2016)

2.6.3. NLTK

NLTK es un framework de cifrado libre para el procesamiento del lenguaje natural que proporciona los instrumentos necesarios para el procesamiento de textos (Loper & Bird, 2002). Esta librería contiene un kit de herramientas que permiten realizar varios procesos de NLP gracias a la ayuda de los siguientes componentes:

- Interprete de palabras frecuentes.
- Segmentador de léxicos y separativos oracionales.
- Lector de corpus para el corpus compilado individualmente para la ejecución del sistema.
- Etiquetador sintáctico para determinar a cada palabra su coste.
- Clasificador bayesiano ingenuo, empleado para el análisis de sentimientos.

2.6.4. Ingeniería de modelos

2.6.4.1. El desarrollo de software dirigido por modelos (MDD)

El Desarrollo de Software Dirigido por Modelos MDD (por sus siglas en inglés: Model Driven software Development) se ha convertido en un nuevo paradigma de

desarrollo software (Pons, Giandini, & Pérez, 2010). MDD promete mejorar el proceso de construcción de software basándose en un proceso guiado por modelos y soportado por potentes herramientas.

El adjetivo “dirigido” (driven) en MDD, a diferencia de “basado” (based), enfatiza que este paradigma asigna a los modelos un rol central y activo: son al menos tan importantes como el código fuente.

La figura 7 muestra la parte del proceso de desarrollo de software en donde la intervención humana es reemplazada por herramientas automáticas. Los modelos pasan de ser entidades contemplativas (es decir, artefactos que son interpretadas por los diseñadores y programadores) para convertirse en entidades productivas a partir de las cuales se derivala implementación en forma automática.

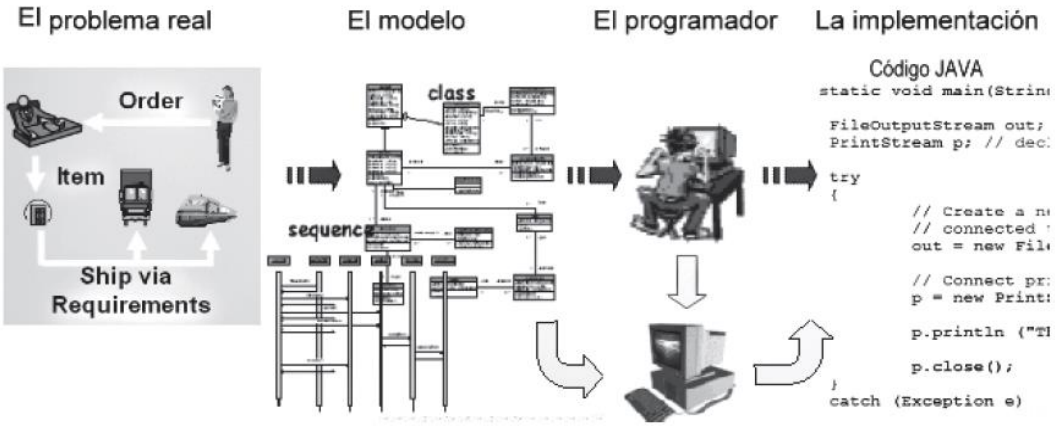

Figura 6. Desarrollo de software dirigido por modelos
Fuente: (Pons, Giandini, & Pérez, 2010)

2.7. Resumen del software

Tabla 9
Detalle de Software

Software	Productor	Versión
Python	Python Software Foundation	3.4.4
Tweepy	Joshua Roesslein	3.5
Matplotlib	John Hunter	2.0.2
Textblob	Steven Loria	0.12

Fuente: (El Autor, 2017)

CAPÍTULO 3

3.1 Descripción de requisitos.

Tabla 10
Descripción de requisitos del sistema.

Requisito	Descripción
Obtención de información escrita	Se requiere extraer la información de la población plasmada en la red social Twitter de forma automática y también en base a una encuesta basada en el test psicológico de Sacks.
Acondicionamiento de la información	Se necesita eliminar los signos de puntuación y palabras vacías del corpus y normalizar el texto para su posterior análisis.
Clasificación de la información	Se necesita determinar si el contenido de un texto se relaciona con situaciones que generan estrés mediante el uso de un algoritmo de clasificación.
Obtención del nivel de estrés en una población	Se requiere que el sistema mida la cantidad de mensajes negativos relacionados con el estrés.

Fuente: (El Autor, 2017)

3.2. Descripción general del sistema

El sistema desarrollado mide el nivel de estrés de cualquier texto escrito y posee dos modos de funcionamiento automático y manual. Para el modo automático se debe primero establecer una conexión con la API de Twitter para extraer el texto escrito por la población en esta red social de acuerdo a varios parámetros de búsqueda. Para el modo manual se necesita contestar un test psicológico de frases incompletas denominado test de Sacks en donde el usuario tiene la libertad de completar cada ítem como desee y al final conocer su nivel de estrés.

Para la creación de los algoritmos y diseño de la interfaz gráfica del usuario se utilizó el lenguaje de programación Python por su versatilidad y disponibilidad de librerías de NLP.

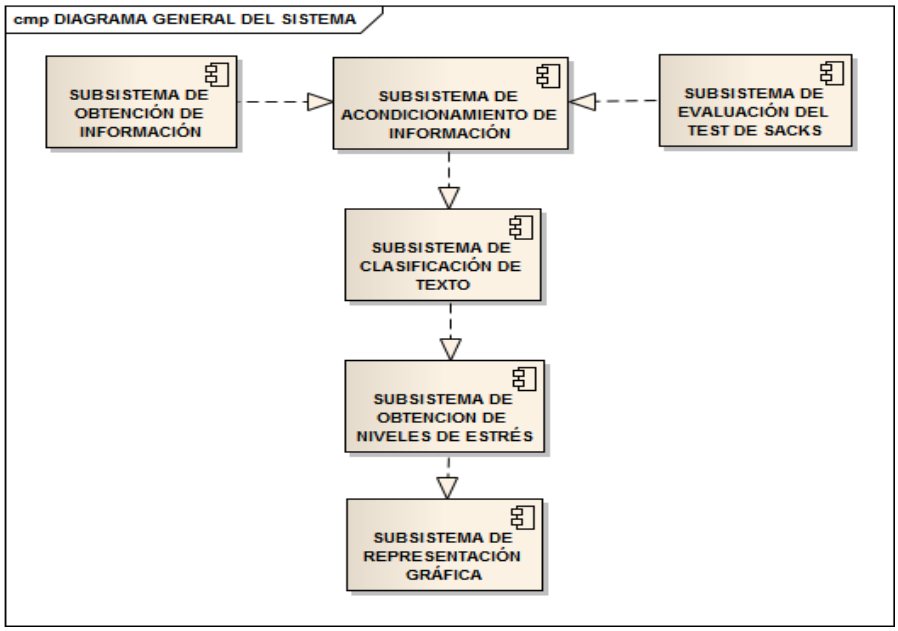

Figura 7 Diagrama de componentes general del sistema

3.3. Descripción del algoritmo general del sistema

Para la implementación del proyecto se realizan los procesos de obtención del texto, procesamiento y limpieza del texto, clasificación, cálculo del nivel de estrés y representación gráfica de los valores medidos.

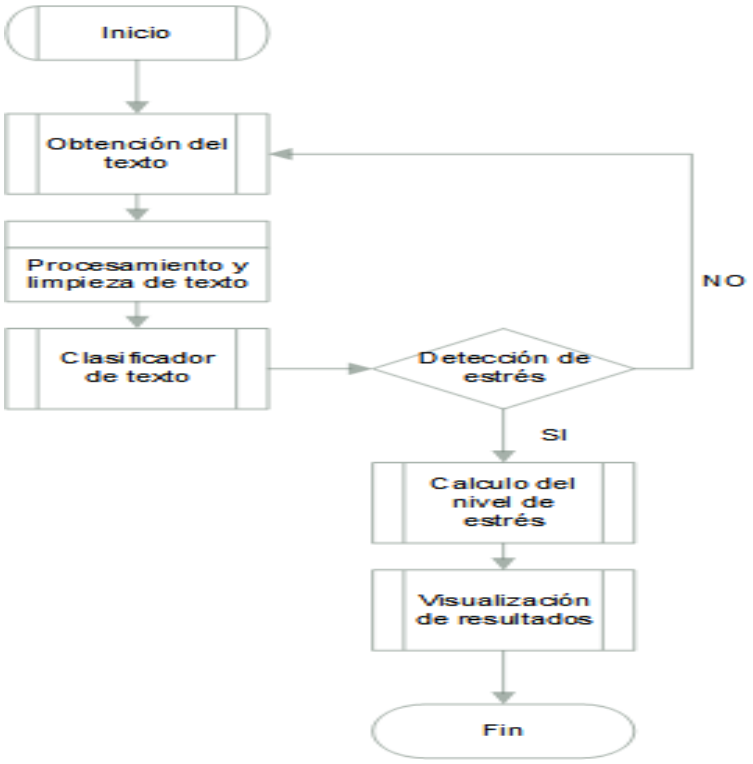

Figura 8 Diagrama del algoritmo general del sistema

3.4. Descripción del subsistema de extracción de información

El texto a ser analizado se obtiene de dos fuentes diferentes la una proviene de la extracción automática de un conjunto de mensajes escritos en la red social Twitter y la otra mediante la evaluación del test psicológico de Sacks que contiene 60 frases a ser completadas por el usuario de forma manual.

3.4.1. Proceso de extracción de información de la red social Twitter

Primero es necesario realizar un proceso de autorización y autentificación de usuario para poder establecer la comunicación desde el sistema hacia el servidor de Twitter.

La función on_data() que es propia de la API de Twitter se encarga de extraer todos los datos de cada Tweet ye los almacena en forma de texto en un objeto de tipo JSON.

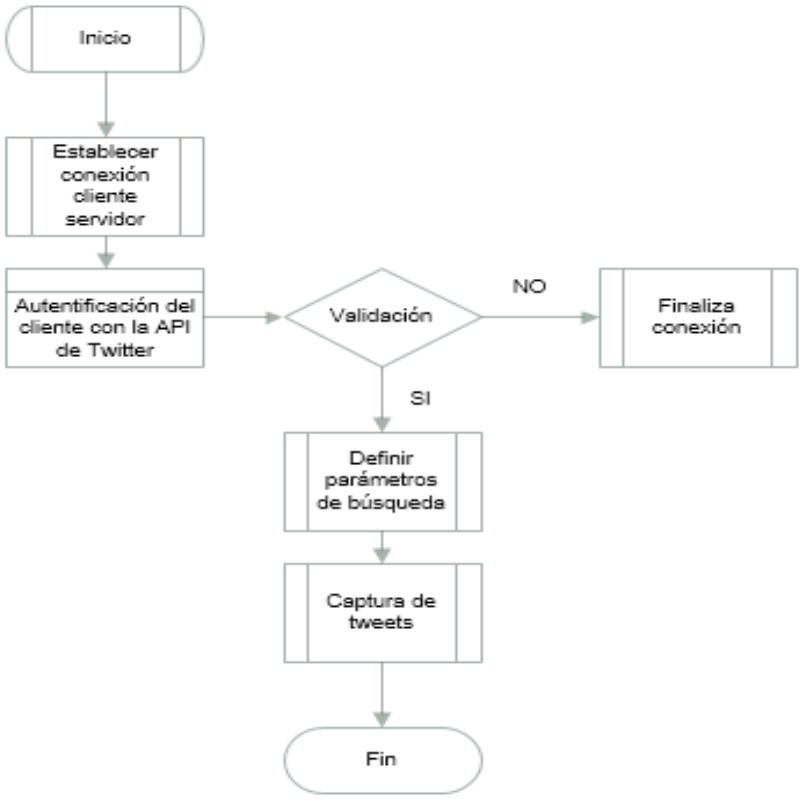

Figura 9 Diagrama del proceso de extracción de Tweets

Se utiliza una función llamada filter() perteneciente a la librería Tweepy para poder realizar la búsqueda de Tweets de acuerdo a la región geográfica. El módulo de

extracción de tweets usa tres Hilos que permiten leer y extraer continuamente los mensajes publicados en Twitter.

3.4.2. Proceso de extracción de información mediante evaluaciones escritas del test de Sacks

Se crea un arreglo para almacenar las 60 frases que deberán ser completadas por el evaluado, las frases aparecen desordenadas en el test original y se debe mantener este esquema ya que es una metodología que ha sido propuesta por el autor del test.

3.5. Descripción del subsistema de acondicionamiento de la información

Se crea una lista de stopwords para el idioma español de Ecuador con todas las palabras irrelevantes y que no aportan mucho al sentido de la oración como los pronombres, artículos, proposiciones, signos de puntuación, y cualquier tipo de símbolo desconocido.

El texto extraído pasa por un proceso de filtrado que se encarga de eliminar todas las palabras vacías y convertir los caracteres a minúsculas para tener un texto normalizado.

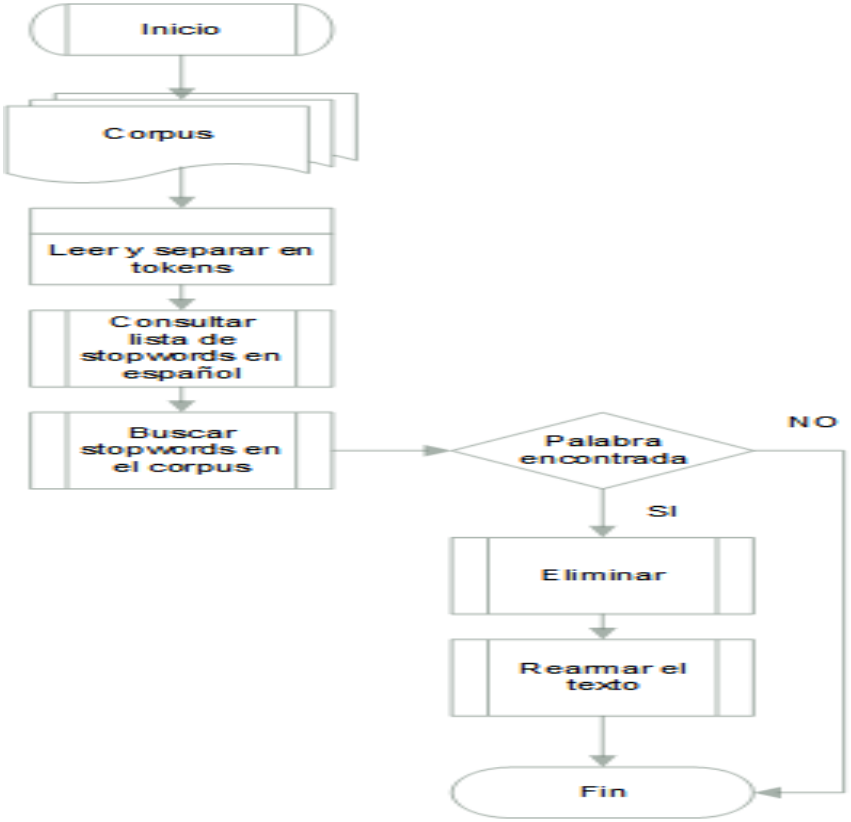

Figura 10 Diagrama de acondicionamiento de la información

3.6. Descripción del subsistema de clasificación de texto

Para este proceso se utiliza un clasificador de texto supervisado Naive Bayes, este clasificador necesita ser entrenado con datos conocidos de acuerdo al requerimiento que en este caso es la detección del estrés.

El propósito del sistema es determinar si el texto escrito por una persona de manera inconsciente refleja estrés por lo que se debe alimentar al entrenador con oraciones positivas y negativas para que el clasificador pueda aprender y categorizarlas como positivas ‘cuando no existe estrés’ y como negativas ‘cuando si existe estrés’, para ello calcula las probabilidades de ocurrencia de acuerdo a los datos conocidos que son ingresados por el programador con anterioridad.

El entrenador es un arreglo que contiene una lista de palabras y frases tanto positivas como negativas. La lista de palabras y frases negativas en su gran mayoría se relacionan a situaciones de estrés, de esta manera el clasificador aprende como clasificar cada texto extraído dese dos fuentes de información: desde la red social Twitter y de las frases del test de Sacks completadas por una persona de un determinado género y edad.

Tabla 11
Ejemplo de palabras y frases introducidas en el entrenador

Número	Oraciones	Clasificación manual
1	a gusto	POS
2	mejor amigo	POS
3	amor	POS
4	miedo	NEG
5	maravillosa	POS
6	buena relación	POS
7	hacer daño	NEG
8	fracasar	NEG
9	tristeza	NEG
10	rabia	NEG

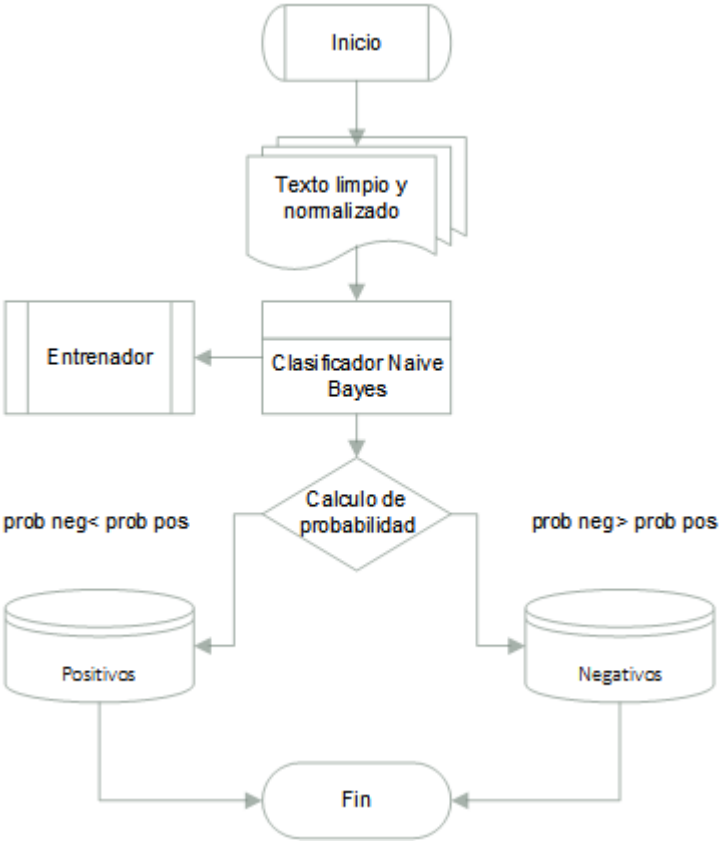

Figura 11 Diagrama del algoritmo de clasificación

3.7. Descripción del módulo de obtención de los niveles de estrés

Se utilizan las funciones de la librería Matplotlib para realizar las operaciones matemáticas, cálculos de porcentajes y para representar las mediciones de los niveles de estrés por medio de gráficas estadísticas de barras y pasteles.

Se calcula el porcentaje de negatividad de acuerdo a la cantidad de mensajes positivos y negativos censados por dos contadores. Se consideran como mensajes negativos todos aquellos que contengan textos relacionados con las fuentes de estrés o estresores.

Los resultados del análisis del test de Sacks se muestran en cuatro gráficas de pastel que indican los porcentajes de positividad y negatividad total obtenidos al realizar el análisis del corpus de cada área del test en forma individual.

Los resultados del análisis de los mensajes escritos extraídos de la red social Twitter se muestran en tres gráficas de barras.

3.8. Diseño de la interfaz gráfica del módulo de extracción de Tweets

La interfaz gráfica de usuario del módulo de extracción se dividirá en seis áreas en donde se ubicarán los componentes que permiten la lectura y escritura de datos y la comunicación con los subsistemas para el procesamiento de la información.

3.8.1. Área 1 de la interfaz basada en recolección de Tweets

En esta área se ubicaron tres botones: El botón Censar, para inicializar los Hilos que permiten al subsistema de extracción de información leer de forma continua los mensajes escritos en el Twitter. El botón Graficar, que se encarga de invocar al subsistema de representación gráfica para desplegar las mediciones de los niveles de estrés mediante una gráfica estadística de barras. El botón Ayuda que despliega una guía de uso del sistema.

3.8.2. Área 2 de la interfaz basada en recolección de Tweets

En esta área se ubicaron dos cajas de texto para el ingreso del tiempo de censado y ejecución de los Hilos.

3.8.3. Área 3 de la interfaz basada en recolección de Tweets

En esta área se ubicaron tres etiquetas que muestran en forma numérica los porcentajes de mensajes negativos y positivos para cada región geográfica: Costa, Sierra y Oriente.

3.8.4. Área 4, 5 y 6 de la interfaz basada en recolección de Tweets

En esta área se ubicaron los resultados de la medición de los niveles de estrés en forma gráfica invocando al subsistema de representación gráfica y al subsistema de obtención de los niveles de estrés.

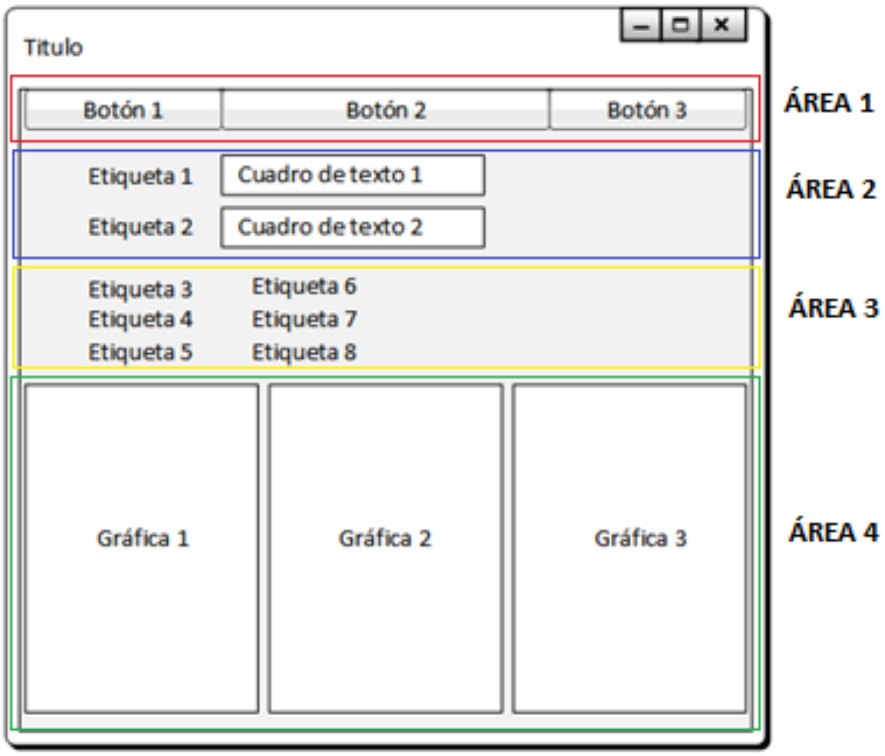

Figura 12. Áreas de la ventana de extracción de Tweets

3.9. Diseño de la interfaz gráfica por el método del test de Sacks

La interfaz gráfica de usuario del módulo para el test de Sacks se dividirá en tres áreas en donde se ubicarán los componentes que permiten la lectura y escritura de datos y la comunicación con los subsistemas para el procesamiento de la información.

3.9.1. Área 1 de la interfaz basada en el test de Sacks

En esta área se ubicaron una Etiqueta 1 y una caja de texto 1 para la recolección de información escrita de las respuestas a las 60 frases del test de Sacks y permite el envió de los datos escritos al subsistema de clasificación de textos y de obtención de los niveles de estrés.

3.9.2. Área 2 de la interfaz basada en el test de Sacks

En esta área se ubicaron los botones 1 y 2 que permiten invocar al subsistema de representación gráfica para la obtención de las cuatro gráficas con porcentajes de oraciones positivas y negativas para cada área del test de Sacks.

3.9.3. Área 3 de la interfaz basada en el test de Sacks

En esta área se realizará la representación gráfica de los valores calculados en el subsistema de obtención de los niveles de estrés y de representación gráfica de usuario.

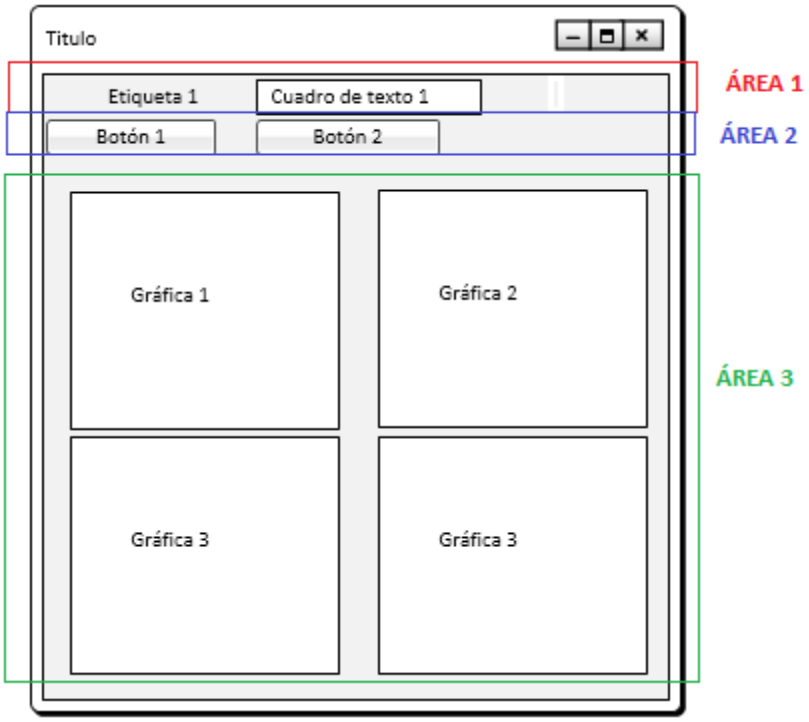

Figura 13. Áreas de la ventana del test de Sacks

CAPÍTULO 4

4.1. Introducción

En este capítulo, se desarrollan distintos escenarios de prueba para evaluar el desempeño del sistema de detección de estrés en textos escritos. El sistema desarrollado obtiene la información de dos fuentes diferentes:

La primera proviene de la red social Twitter de donde se extraen un número finito de Tweets de las 3 regiones del Ecuador en un periodo de tiempo establecido por el usuario.

La segunda proviene de un conjunto de 60 frases incompletas del test psicológico de Sacks que fueron completadas por una persona de un determinado género y edad.

Previamente se entrenó al clasificador mediante la inicialización de un arreglo que contiene una lista de palabras y frases pequeñas tanto positivas como negativas. En el caso de las palabras y frases negativas la mayoría están relacionadas con las fuentes de estrés o estresores por ser este el tema principal de análisis.

4.2. Evaluación del sistema

Figura 14 Pantalla principal del sistema

4.3. Escenario de prueba 1

Se realizó una prueba de clasificación para 15 textos de Tweets extraídos de las regiones Sierra Costa y Oriente y se compraron los resultados del algoritmo con los de la clasificación manual y se obtuvieron los casos de acierto VN (Verdaderos Negativos), VP (Verdaderos Positivos) y de error FN (Falsos Negativos), FP (Falsos Positivos) para realizar la evaluación del clasificador. Se utilizaron 3 métricas el accuracy, la precisión, y el recall.

4.3.1. Clasificación de textos obtenidos desde Twitter

Tabla 12
Clasificación de tweets región Sierra, Costa y Oriente

Clasificación de Tweets para la región Sierra				
Número	**Tweet**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
1	la aventura es tu mejor compañera en baños de agua santa déjate seducir por los deportes extremos	Positivo	Positivo	VP
2	me pasa algo curioso y es que todos los días mis amigas se ríen de mi	Negativo	Positivo	FP
3	basta de perseguir a quienes denuncian la corrupción	Negativo	Negativo	VN
4	el remedio peor que la enfermedad	Negativo	Negativo	VN
5	tras las rejas dos presuntos autores de robo de vehículo en ibarra operativo realizado por gom imbabura	Positivo	Negativo	FN

Continua ⟶

Clasificación de Tweets para la región Costa				
Número	**Tweet**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
6	en el asalto a bus san luis los delincuentes asesinaron de un disparo al conductor oriundo del cantón sta isabel	Negativo	Negativo	VN
7	Esmeraldas da miedo jajaja lo que se inventa la gente	Positivo	Negativo	FN
8	Que hermoso país donde ser roban el oro a manos llenas	Negativo	Positivo	FP
9	se inaugura uem simón plata torres en esmeraldas educación de calidad derecho de todos	Positivo	Positivo	VP
10	porque la única obligación del hombre es ser feliz y hacer feliz a alguien	Positivo	Positivo	VP
Clasificación de Tweets para la región Oriente				
Número	**Tweet**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
11	el futuro de la patria es la juventud zamorachichipe contigo	Positivo	Positivo	VP
12	La furia de las feministas suele aparecer cuando les tocan su becerro de oro	Negativo	Positivo	FP
13	turismo arte cultura diversidad organización resaltaron en festival culturas d gadmpastaza y confeniae	Positivo	Positivo	VP

Continua ⟶

14	pj de zamoraec afortunadamente logra detener a tres personas que intentaban robar un taxi de la localidad	Negativo	Negativo	VN
15	niños de varios sectores del cantón macas, en la provincia de morona santiago se divierten con las colonias vacacionales del buen vivir	Positivo	Positivo	VP

4.3.2. Cálculo de las métricas para el método de extracción de Tweets

Tabla 13
Métricas para el método de extracción de Tweets

Datos	VN=4, VP=6, FP=3, FN=2
Accuracy	66.6%
PrecisiónPOS	66.6%
RecallPOS	75%
PrecisiónNEG	66.6%
RecallNEG	57,1%

4.3.3. Cálculo de la efectividad método de extracción de Tweets.

Tabla 14.
Efectividad del clasificador el método de extracción de Tweets

Métrica	**Efectividad del clasificador**
$\mathbf{F1_{POS}}$	70.5%
$\mathbf{F1_{NEG}}$	61.4%

De acuerdo a los resultados se concluye que existe mayor efectividad del clasificador en clasificar Tweets positivos, pero en ninguno de los casos se alcanza el 100% debido que existen Tweets que son clasificados de forma incorrecta de acuerdo al valor de

Accuracy solo el 66.6% de Tweets fueron clasificados de forma correcta el 33.4% restante de Tweets fueron clasificados de forma incorrecta.

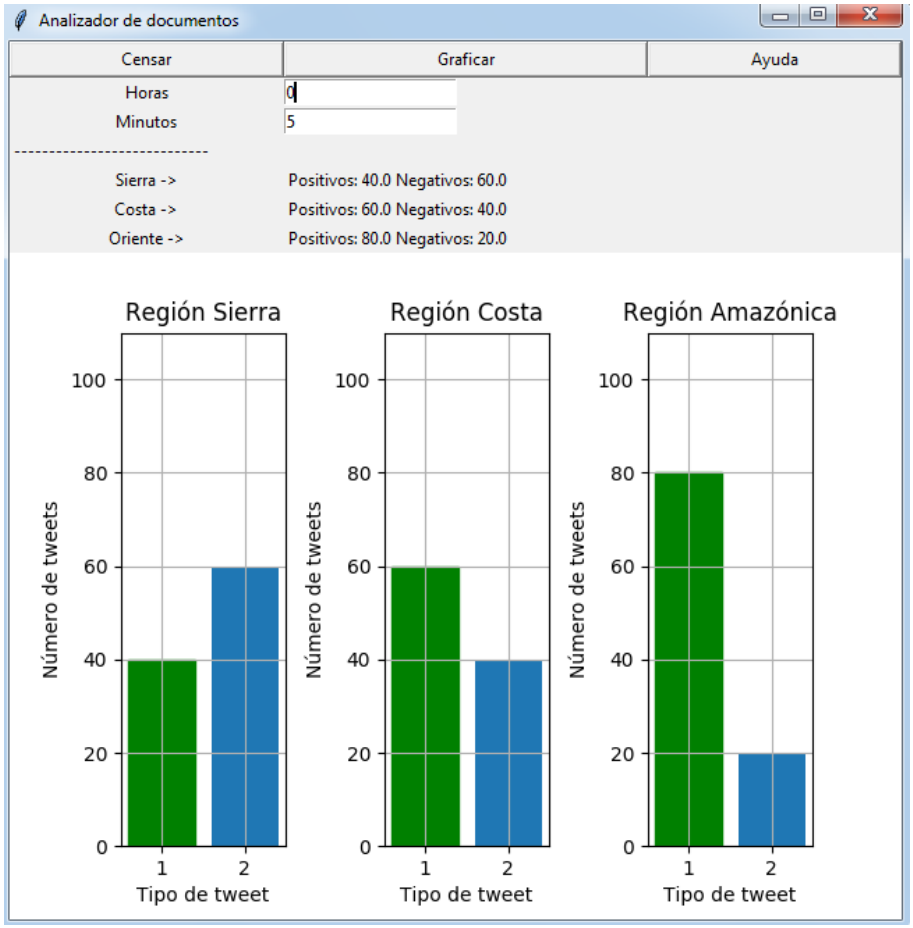

Figura 15 Gráfica de niveles de estrés por regiones del Ecuador

Tabla 15
Porcentajes de Tweets positivos y negativos

Clasificación	Porcentajes Sierra	Porcentajes Costa	Porcentajes Oriente
Positivos	40%	60%	80%
Negativos	60%	40%	20%

De acuerdo a los resultados se concluye que existe mayor estrés en la región Sierra debido a que en el tiempo de censado de 5 minutos se encontró un mayor número de Tweets con textos negativos relacionados con el estrés.

4.4. Escenario de prueba 2

Se realizó evaluación del sistema basado en el test de Sacks que consta de 60 frases incompletas obtenidas de la aplicación del test a una persona anónima de 25 años de edad y de género masculino.

4.4.1. Clasificación de frases del área de adaptación familiar

Tabla 16
Clasificación de frases de actitud hacia el padre

Frases de actitud hacia el padre				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
1	Siento que mi padre raras veces se enoja	POS	POS	VP
16	Si mi padre tan sólo siempre siga vivo	POS	POS	VP
31	Desearía que mi padre nunca muera	POS	NEG	FN
46	Siento que mi padre es perfecto	POS	POS	VP

Tabla 17
Clasificación de frases de actitud hacia la madre

Frases de actitud hacia la madre				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
14	Mi madre la mejor de todas	POS	POS	VP
29	Mi madre y yo siempre juntos	POS	POS	VP

Continua ⟶

44	Creo que la mayoría de las madres son cariñosas	POS	POS	VP
59	Me agrada mi madre, pero no hay pero	POS	POS	VP

Tabla 18
Clasificación de frases de actitud hacia la unidad familiar

Frases de actitud hacia la unidad familiar				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
12	Comparada con las demás familias, la mía es extraordinaria	POS	POS	VP
27	Mi familia me trata como un niño	NEG	POS	FP
42	La mayoría de las familias que conozco son buenas	POS	POS	VP
57	Cuando era niño (a), mi familia me cuidaba mucho	POS	POS	VP

4.4.2. Clasificación de frases del área sexual

Tabla 19
Clasificación de frases de actitud hacia los hombres/las mujeres

Frases de actitud hacia los hombres/las mujeres				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
10	Mi idea de mujer (hombre) perfecta (o) nunca comete errores	POS	NEG	FN

Continua ⟶

25	Pienso que la mayoría de las muchachas (os) son buenos	POS	POS	VP
40	Creo que la mayoría de las mujeres (hombres) son bonitas	POS	POS	VP
55	Lo que menos me gusta de las mujeres (hombres) que se hagan las víctimas	NEG	NEG	VN

Tabla 20
Clasificación de frases de actitud hacia las relaciones heterosexuales

Frases de actitud hacia las relaciones heterosexuales				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
11	Cuando veo un hombre y a una mujer juntos son felices	POS	POS	VP
26	Yo creo que la vida matrimonial es perfecta	POS	POS	VP
41	Si tuviera relaciones sexuales disfrutaría mucho	POS	POS	VP
56	Mi vida sexual activa	POS	POS	VP

4.4.3. Clasificación de frases del área de relaciones interpersonales

Tabla 21
Clasificación de frases de actitud hacia amigos y conocidos

Frases de actitud hacia amigos y conocidos				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
8	Creo que es un verdadero amigo el que está en las buenas y malas	POS	POS	VP
23	No me gusta la gente con baja autoestima	NEG	NEG	VN
38	La gente que más me agrada mi familia	POS	POS	VP
53	Cuando no estoy, mis amigos tampoco	NEG	POS	FP

Tabla 22
Clasificación de frases de actitud hacia colegas en el trabajo

Frases de actitud hacia colegas en el trabajo o escuela				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
13	En las labores me llevo mejor con todos	POS	POS	VP
28	Aquellos con los que trabajo son excelentes	POS	POS	VP
43	Me gusta trabajar con la gente que trabaje	POS	NEG	FN
58	La gente que trabaja conmigo, generalmente se sienten a gusto	POS	POS	VP

Tabla 23
Clasificación de frases de actitud hacia superiores en el trabajo

Frases de actitud hacia superiores en el trabajo o escuela				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
6	Las personas que están sobre mí son mis padres	POS	POS	VP
21	En la escuela, mis maestros eran comprensivos	POS	POS	VP
36	Cuando veo el jefe venir me asusta	NEG	NEG	VN
51	La gente a quien yo considero mis superiores mi familia	POS	POS	VP

Tabla 24
Clasificación de frases de actitud hacia los subordinados

Frases de actitud hacia los subordinados				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
4	Si yo estuviera a cargo todos participan	POS	POS	VP
19	Si la gente trabaja para mí para que se sientan bien	POS	POS	VP
34	La gente que trabaja para mí son felices	POS	POS	VP
49	Lo que más deseo en la vida es tener hijos	POS	POS	VP

4.4.4. Clasificación de frases del área de autoconcepto

Tabla 25
Clasificación de frases de actitud hacia los temores

Frases de actitud hacia los temores				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
7	Sé que es tonto, pero tengo miedo de fracasar	NEG	NEG	VN
22	La mayoría de mis amistades no saben que tengo miedo de las arañas	NEG	NEG	VN
37	Quisiera perder el miedo de fracasar	POS	NEG	FN
52	Mis temores en ocasiones me obligan a cometer errores	NEG	POS	FP

Tabla 26
Clasificación de frases de actitud hacia los sentimientos de culpa

Frases de actitud hacia los sentimientos de culpa				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
15	Haría cualquier cosa por olvidar la vez que me choque	NEG	POS	FP
30	Mi más grande error fue tomar alcohol	NEG	POS	FP

Continua ⟶

45	Cuando era más joven me sentía culpable de no hablar	NEG	POS	FP
60	La peor cosa que he hecho no recuerdo	NEG	NEG	VN

Tabla 27
Clasificación de frases de actitud hacia las metas

Frases de actitud hacia las metas				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
3	Siempre anhelé ser astronauta	POS	POS	VP
18	Sería perfectamente feliz si termino mis estudios	POS	POS	VP
33	Mi ambición secreta en la vida llegar a tener éxito	POS	POS	VP
48	Cuando doy órdenes, yo soy tranquilo	POS	POS	VP

Tabla 28
Clasificación de frases de actitud hacia las propias capacidades

Frases de actitud hacia las propias capacidades				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
2	Cuando tengo mala suerte me quejo mucho	NEG	NEG	VN
17	Siento que tengo la habilidad para jugar	POS	POS	VP

Continua ⟶

32	Mi mayor debilidad es ser sentimental	NEG	NEG	VN
47	Cuando la suerte se vuelve en contra mía eso no quiero	NEG	POS	FP

Tabla 29.
Clasificación de frases de actitud hacia el pasado

Frases de actitud hacia el pasado				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
9	Cuando era niño (a) jugaba mucho	POS	POS	VP
24	Antes tenía muchos sueños	NEG	POS	FP
39	Si fuera joven otra vez voy al colegio	POS	POS	VP
54	Mi más vivido recuerdo de la infancia no recuerdo	NEG	POS	FP

Tabla 30
Clasificación de frases del área de actitud hacia el futuro

Frases de actitud hacia el futuro				
Número	**Frase completada**	**Clasificación manual**	**Clasificación del algoritmo**	**Caso**
5	El futuro me parece interesante	POS	POS	VP
20	Yo espero tener hijos	POS	POS	VP

Continua ⟶

35	Algún día yo seré un gran profesional	POS	POS	VP
50	Dentro de algún tiempo seré papá	POS	POS	VP

4.4.5 Cálculo de las métricas para el método de extracción de información mediante la evaluación del test de Sacks.

Tabla 31
Métricas para el método de evaluación del test de Sacks

Datos	VN=8, VP=39, FP=9, FN=4
Accuracy	78.3%
PrecisiónPOS	81.2%
RecallPOS	90.7%
PrecisiónNEG	66.6%
RecallNEG	47%

4.4.6 Cálculo de la efectividad para el método de evaluación del test de Sacks.

Tabla 32
Efectividad del clasificador para el método de evaluación del test de Sacks

Métrica	**Efectividad del clasificador**
$\boldsymbol{F1_{POS}}$	85.6%
$\boldsymbol{F1_{NEG}}$	55.1%

De acuerdo a los resultados se concluye que existe mayor efectividad del clasificador en clasificar frases positivas, pero en ninguno de los casos se alcanza el 100% debido que existen frases que son clasificados de forma incorrecta de acuerdo al valor de Accuracy solo el 66.6% de frases fueron clasificados de forma correcta el 33.4% restante de frases fueron clasificados de forma incorrecta.

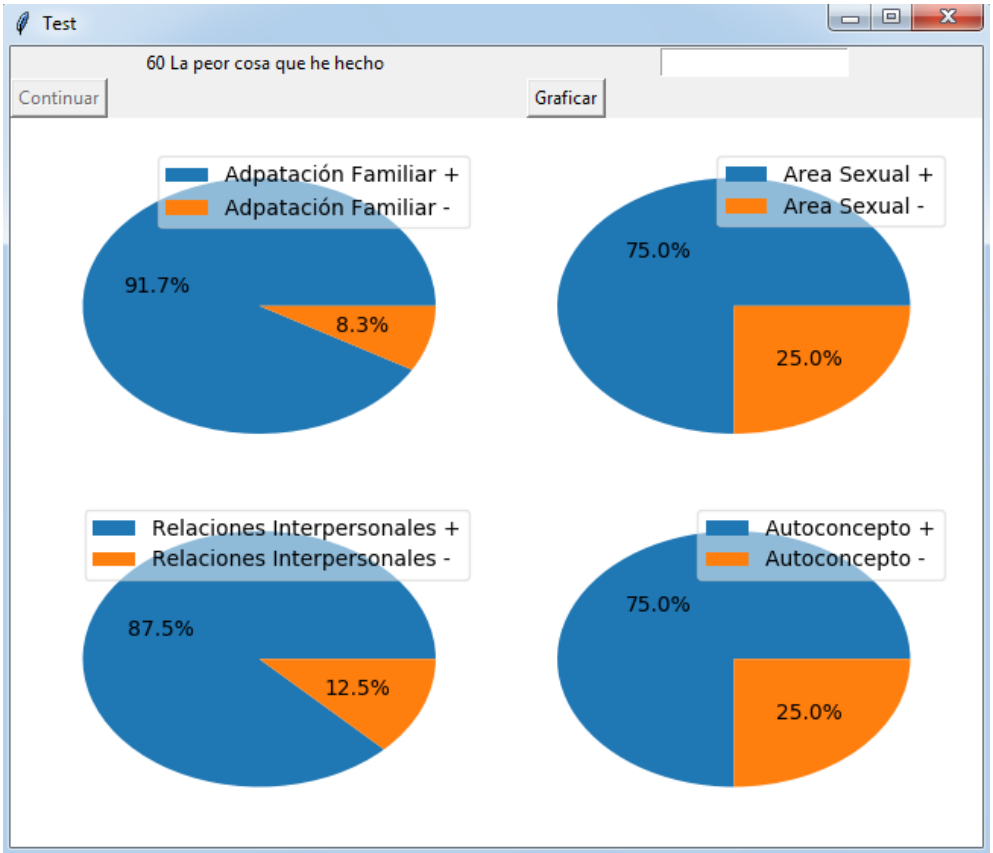

Figura 16 Gráfica de resultados por área del test de Sacks

Tabla 33.
Porcentajes de frases positivas y negativas del test de Sacks evaluado

Porcentaje	Área de adaptación familiar	Área sexual	Área de relaciones interpersonales	Área de autoconcepto
Positivo	91.7%	75%	87.5%	75%
Negativo	8.3%	25%	12.5%	25%

Estos resultados indican que la persona no posee conflictos en ninguna de las cuatro áreas de estudio del test de Sacks ya que los porcentajes de frases completadas de forma positiva son mayores a las frases completadas de forma negativa.

CAPÍTULO 5

5.1. Conclusiones

- Se realizó la limpieza y normalización del corpus mediante procesos de NLP tokenización, conversión a letra minúscula, búsqueda y comparación de textos, para obtener un texto sin palabras vacías, números, ni signos de puntuación que no aportan mucho al sentido de una oración, esto evitó tener oraciones repetitivas en el entrenador y reducir el tiempo de procesamiento del sistema de medición del estrés en textos escritos, también se utilizaron los mismos procesos de NLP para la recuperación de información que es obtenida a través de dos fuentes: desde la red social Twitter y a partir de la evaluación del test psicológico de Sacks a un grupo de 30 personas de entre 25 y 30 años de edad de género masculino y femenino.
- Se utilizó un clasificador tipo Naive Bayes que contiene un algoritmo de aprendizaje supervisado y que fue previamente entrenado con toda la información disponible acerca del estrés, en este entrenador se ingresaron palabras y frases positivas y negativas escritas en el idioma español de Ecuador y clasificadas de forma manual, de esta manera se pudo detectar y medir los niveles de estrés en el discurso escrito de una persona.
- La metodología de extracción de información a partir de la evaluación del Test de Sacks permitió conocer de forma más detallada las actitudes de la persona evaluada y determinar las áreas de afección, mediante el cálculo de los porcentajes de positivad y negatividad en las respuestas. Debido a que el test está diseñado para no limitar a que el usuario responda con palabras simples como SI o NO la mayoría de frases son completadas con información relevante que permite al clasificador tener resultados más exactos.
- El test de Sacks permitió extraer información relevante de cuatro áreas: área de adaptación familiar, área sexual, área de relaciones interpersonales, y área de autoconcepto, dando libertad a la persona de completar como crea conveniente cada una de las 60 frases que componen el test, la metodología propuesta puede ser utilizada para detectar otros tipos de problemas que afecten a la persona evaluada a parte del estrés.

- La metodología utilizada para la extracción de los mensajes escritos en la red social Twitter por medio de la API Tweepy permitió configurar ciertos parámetros de búsqueda avanzados para así cumplir con los propósitos del sistema, siendo uno de ellos filtrar los tweets por regiones del Ecuador mediante geolocalización para lo cual fue necesario conocer las coordenadas exactas de Latitud y Longitud geográfica para establecer un área de búsqueda.
- Se obtuvieron porcentajes de efectividad positiva de 70,5 y 85.6 % y de efectividad negativa de 61.4% y 55.1% para los dos métodos de extracción de información respectivamente, en ninguno de los casos se alcanzó el 100% de efectividad debido a que para la obtención de resultados el sistema utiliza un clasificador supervisado el cual aplica un método probabilístico, con un entrenador que contiene información proporciona por el programador y en ocasiones cuando aparece un texto que se relaciona poco o nada con la información que posee el entrenador, el sistema clasifica el texto de forma incorrecta apareciendo dos casos de error llamados falsos positivos FP y falsos negativos FN, reduciendo así el porcentaje de efectividad del clasificador.

5.2. Recomendaciones

- Para evitar conflictos con las librerías de Python se debe verificar las versiones del software que ya son estables y procurar trabajar con la última versión disponible en su página de forma gratuita.
- Para la extracción de tweets de varias regiones geográficas es recomendable usar una cuenta de usuario diferente de Twitter por región debido a que el canal de la API Tweepy no permite transmitir varios flujos de información al mismo tiempo.
- El entrenador debe poseer una cantidad moderada de información que permita al clasificador categorizar correctamente los mensajes ya que si se satura de información aumenta la probabilidad de error aumentado la cantidad de falsos positivos y falsos negativos.

5.3. Trabajos futuros

- Se propone la investigación de técnica y métodos que permitan determinar el tamaño ideal de la muestra de información de una población, enfocado a la minería de opinión y recuperación de información en redes sociales para obtener la mayor cantidad de información confiable y obtener mejores resultados en el procesamiento de los datos.
- Se propone montar el sistema desarrollado en un servidor con libre acceso que permita conocer las afectaciones de la población ecuatoriana y poder realizar la evaluación del test de Sacks a cualquier persona que lo desee.
- Se propone el desarrollo de más herramientas de NLP para el idioma español de Ecuador ya que en la actualidad es una limitación debido a que la gran mayoría de herramientas de software son desarrolladas para el idioma ingles y las pocas disponible para el idioma español son de pago y muy básicas.

ANEXOS

1. Anexos Visuales

Se encuentran en el CD en la carpeta videos, son dos videos con la explicación del funcionamiento y la forma de uso del sistema desarrollado.

2. Otros Anexos

Se encuentran en el CD en la carpeta escrito, que contiene el manual de usuario del sistema.

REFERENCIAS

Angeli, G., Manning, C. D., & Jurafsky, D. (2012). Parsing time: Learning to interpret time expressions. *In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 446-455.

Arrington, M. (2006). Odeo releases twttr. *TechCrunch*.

Baeza Yates , R., & Ribeiro Neto, B. (1999). Modern information retrieval (Vol. 463). *New York: ACM press*.

Bird, S. (2006). NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions (pp. 69-72). *Association for Computational Linguistics*.

Blanco Rodríguez , M. M., Cabrera Hernández, R., Arrieta Vergara, K., & M. A. (2015). Relación entre satisfacción vocacional, estilo de afrontamiento y estrés percibido con el Síndrome de Burnout en estudiantes de odontología (Doctoral dissertation, Universidad de Cartagena).

Breilh, J. (1992). Trabajo hospitalario, estrés y sufrimiento mental: deterioro de la salud de los internos en Quito, Ecuador.

Cassaretto, M., Chau, C., Oblitas, H., & Valdez, N. (2003). Estrés y afrontamiento en estudiantes de psicología. *Revista de psicología*, 21(2), 363-392.

D'Monte, L. (2017, enero 15). *Swine flu's tweet tweet causes online flutter*. Retrieved from http://www.business-standard.com/india/news/swine-flu%5Cs-tweet-tweet-causes-online-flutter/356604/

Donoso, B. P., Herrera, M. D., & Aguinaga, G. (2004). La Promoción de Salud en el ECUADOR.

Duval, F., González, F., & Rabia, H. (2010). Neurobiología del estrés. Revista chilena de neuro-psiquiatría. 48(4), 307-318.

Ekman , P., & Friesen , W. (1978). *Facial Action Coding System:*. Palo Alto: Consulting Psychologists Press.

Fernández,, C. A., Fernández E, M. A., & Pesqueira, G. S. (2000). Problemas semánticos en la adaptación del POMS al castellano. Psicothema. 47-51.

Flores Vivar, J. M. (2009). Nuevos modelos de comunicación perfiles y tendencias en las redes sociales. *Comunicar*.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*.

García, A. (2012). *Inteligencia artificial: fundamentos, práctica y aplicaciones.* RC Libros.

Gelbukh, A. (2010). Procesamiento de lenguaje natural y sus aplicaciones. Korpus Sapiens. *Sociedad Mexicana de inteligencia artificial*, 1.

Guinovart, J. G. (1998). Fundamentos de Lingüística Computacional: bases teóricas, líneas de investigación y aplicaciones. *Bibliodoc: anuari de biblioteconomia, documentació i informació*, 135-146.

Henry, J. D., & Crawford, J. R. (2005). *ritish journal of clinical psychology*, 44(2), 227-239.

Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *British journal of clinical psychology*.

Hernández, J. M., Polo, A., & Pozo, C. (1996). *Inventario de Estrés Académico. Servicio de Publicaciones de la Universidad Autónoma de Madrid,*. Madrid.

Holmes, T. H., & Rahe, R. H. (1967). *The social readjustment rating scale. Journal of psychosomatic research.*

Huaquin, V., Moyano, R., & Loaiza, R. (2000). Construcción de un test para medir estrés general universitario. *Revista Chilena de Psicología*.

Hudlicka, E. (2003). To feel or not to feel: The role of affect in human–computer interaction. *International journal of human-computer studies*, 1-32.

INEC. (2014). Encuesta de Condiciones de Vida 2013-2014.

Izarraga, K., & Serra, J. (2016, diciembre 14). *CÓMO MEDIR EL ESTRÉS?* Retrieved from Bizkaia: www.bizkaia.eus/dokumentuak/04/kirolak/.../Como%20medir%20el%20estres

Joe, C. V. (2014). Developing Tamil emotional speech corpus and evaluating using SVM. *In Science Engineering and Management Research (ICSEMR), 2014 International Conference on IEEE*, 1-6.

Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). Multinomial Naive Bayes for Text Categorization Revisited. *In Australian Conference on Artificial Intelligence*, 488-499.

Laza, R., & Pavón, R. (2010). Clasificador Bayesiano de Documentos MedLine a partir de Datos No Balanceados.

Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. In Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1. *Association for Computational Linguistics.*, 63-70.

Macías, A. B. (2005). Características del estrés académico de los alumnos de educación media superior. *Revista Electrónica PsicologiaCientifica. com*.

Martí, M. A. (2003). *Tecnologías del lenguaje.* Barcelona: Editorial UOC.

Martínez Cámara, E., Valdivia, M., Teresa, M., Perea Ortega, J. M., & Ureña López, L. A. (2011). Técnicas de clasificación de opiniones aplicadas a un corpus en español.

Medina, P. N., Starostenko, O., & Ruiz Castillo, O. (2013). Desarrollo de un avatar animado con. *Komputer Sapiens*, 8-12.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 39-41.

Minsky, M. L. (1967). *Computation: finite and infinite machines.* Prentice-Hall, Inc.

Otero, P. G., & González, M. G. (2000). Técnicas de Procesamiento del Lenguaje Natural en la Recuperación de Información. Centro de Investigación sobre Tecnoloxías da Lingua.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 79-86.

Pérez, A., De Macedo, M., Canelones, P., & Castés, M. (2002). Niveles de inmunoglobulina "A" secretora en condición de estrés académico en estudiantes de medicina. *Revista Electrónica de Motivación y Emoción*.

Perone, C. S. (2009). Pyevolve: a Python open-source framework for genetic algorithms. *Acm Sigevolution*, 12-20.

Picard, R. W., & Picard, R. (1997). *Affective computing.* Cambridge: MIT press.

Pons, C., Giandini, R., & Pérez, G. (2010). *Desarrollo de software dirigido por modelos.* Buenos Aires: EDLULP.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. *In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 275-281.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Rauf Subhani, A., Mumtaz, W., Naufal, M., Kamel, N., & Saeed Malik, A. (2017). Machine Learning Framework for the Detection of Mental Stress at Multiple Levels. *IEEE Access*, 1-11.

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the Association for Information Science and Technology*, 129-146.

Roesslein, J. (2015). tweepy Documentation.

Sacks, J. M., & Levy, S. (1950). The sentence completion test. Projective psychology: Clinical approaches to the total personality. 357-402.

Salton, G., & Buckley, C. (1988). Parallel text search methods. *Communications of the ACM*, 202-215.

Samuel, A. L. (1969). ome studies in machine learning using the game of checkers. *II—Recent progress. Annual Review in Automatic Programming*, 1-3.

Sandín, B. (2003). El estrés: un análisis basado en el papel de los factores sociales. *Revista internacional de psicología clínica y de la salud= International journal of clinical and health psychology*, 141-157.

Sarno, D. (2009). Twitter creator Jack Dorsey illuminates the site's founding document. Part I. *Los Angeles Times*, 18.

Seara, & Robles. (2017). *Escalas de Apreciación del Estrés.* Madrid: TEA.

Selye, H. (1964). *El estrés. La tensión en la vida.* Buenos Aires: General Fabril Editora SA.

Sender, Valles, Puig, Salamero, & Valdés. (2004). *Escala de percepción de estrés.*

Senplades, S. N. (2013, Octubre 2). Plan Nacional del Buen Vivir.

Shi, Z., Rui, H., & Whinston, A. B. (2013). Content sharing in a social broadcasting environment: evidence from twitter.

Slipak, O. E. (1991). ALCMEON 3 Historia y concepto del estrés (1ra. Parte). *Alcmeon*, 355-360.

Strzalkowski, T., Lin, F., Perez Carballo, J., & Wang, J. (1997). Natural language information retrieval TREC-6 report. *In TREC*, 347-366.

Toledo Costa, A., Godoy Guerra, M. T., & Suárez Puente, Z. (2008). *El análisis semántico, sintáctico y pragmático en la enseñanza de los contenidos gramaticales.* VARONA.

Trucco, M. (2002). Estrés y trastornos mentales: aspectos neurobiológicos y psicosociales. *Revista chilena de neuro-psiquiatría*, 40, 8-19.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th annual meeting on association for computational linguistics*, 417-424.

Vilares, J. (2006). Aplicaciones del Procesamiento del Lenguaje Natural en la Recuperación de Información en Español. 36.

Villamil Gómez, W., Alba Silvera, L., Menco Ramos, A., Gonzalez Vergara, A., Molinares Palacios, T., Barrios Corrales, M., & Rodríguez Morales, A. J. (2015). Congenital chikungunya virus infection in Sincelejo. *Colombia: a case series. Journal of tropical pediatrics*, 61(5), 386-392.

Viñas, F., & Caparrós, B. (2000). Afrontamiento del período de exámenes y sintomatología somática autoinformada en un grupo de estudiantes universitarios. *Psicología.com*.

Zhai, J., & Barreto, A. (2006). Stress detection in computer users based on digital signal processing of noninvasive physiological variables. *EMBS'06. 28th Annual International Conference of the IEEE* , 1355-1358.